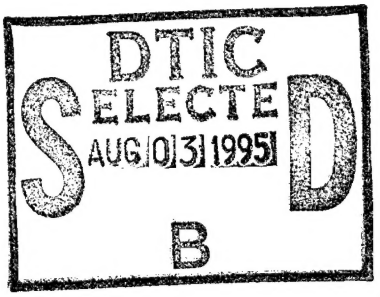
ARMY RESEARCH LABORATORY

Target Acquisition: Human Observer Performance Studies and TARGAC Model Validation

by Dr. J. M. Valeton
Dr. P. Bijl
TNO Human Factors Research Institute
The Netherlands

edited by
Patti Gillespie
Battlefield Environment Directorate



ARL-CR-203 April 1995

19950802 033

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

The citation of trade names and names of manufacturers in this report is not to be construed as official Government indorsement or approval of commercial products or services referenced herein.

Destruction Notice

When this document is no longer needed, destroy it by any method that will prevent disclosure of its contents or reconstruction of the document.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1995	Final

4. TITLE AND SUBTITLE

Target Acquisition: Human Observer Performance Studies and TARGAC Model Validation

5. FUNDING NUMBERS

6. AUTHOR(S)

Dr. J. M. Valeton and Dr. P. Bijl

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TNO Human Factors Research Institute
P.O. Box 23
3769 ZG Soesterberg
The Netherlands

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-CR-203

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783-1145

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARL-CR-203

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

A

13. ABSTRACT *(Maximum 200 words)*

Human target acquisition performance was studied using the thermal imagery that was collected during Battlefield Emissives Sources Trials under the European Theater Weather and Obscurants (BEST TWO), organized by NATO AC243/Panel4/RSG.15 in 1990. Recognition and identification probabilities were measured for a large number of stationary and moving targets at ranges between 1 and 4 km. Target detection was not investigated in a number of carefully controlled laboratory experiments.

The target acquisition model (TARGAC) was validated by comparing its predictions with observed recognition ranges. For all trials used in the observer experiments, TARGAC predictions were calculated on the basis of meteorological, target, background, and time information measured in the field.

The major conclusion of the observer experiments is that the human acquisition performance depends considerably on factors such as target structure, local terrain structure, and cognitive factors. In TARGAC, these factors are not modeled. For the BEST TWO situation, the model predictions were determined solely by target size and thermal imager resolution limit.

A quantitative comparison between actual and predicted recognition ranges shows that TARGAC systematically underestimates human acquisition performance: on average, observed recognition ranges are a factor of 1.8 longer than the model predictions. Further, the model does not make accurate predictions for individual targets on specific backgrounds: the ratio between observed and predicted recognition range varies between 0.9 to 3.6 (95 percent criterion), i.e., a factor of 4. The routine TARGAC uses to calculate electro-optical and human visual system performance is based on the widely-used Night Vision Electro-Optics Sensors Directorate Static Performance Model. Hence, the present findings may also be relevant for the validity of other models.

A number of software errors found in TARGAC are documented; most problems have been fixed.

14. SUBJECT TERMS

target acquistion, target acquisiton models, thermal infrared, observer performance, target recognition, target identification

15. NUMBER OF PAGES

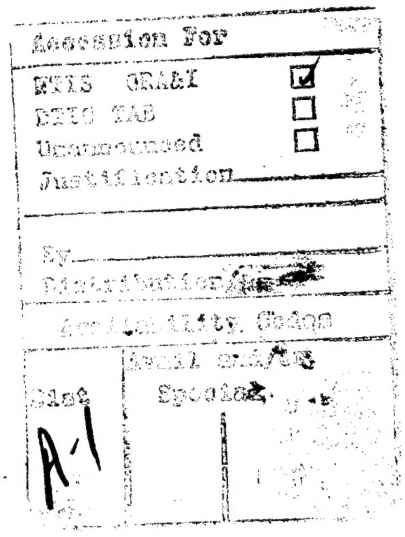
192

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

Acknowledgments

We thank J. Varkevisser, A. Everts, and A. J. C de Reus for expert technical assistance with the development of hardware and software for the observer performance experiments and with the target acquisition (TARGAC) validation analyses. We thank Dr. Patti Gillespie from the U.S. Army Research Laboratory, Battlefield Environment Directorate for help compiling the TARGAC input files, guidance, and scientific support.



Contents

Figures

Appendix Figures

Tables

1. General Overview and Conclusions

1.0 Models and Validation

Target acquisition (TA) models predict how well human observers, using an optical or electro-optical (EO) viewing device, are able to detect, recognize, or identify military targets. The input variables are the properties of the targets and backgrounds, the atmospheric conditions, and the physical properties of the viewing device used. The output is the probability of correct detection, recognition, or identification as a function of target range. TA models are used, for example, in tactical decision aids (TDAs), in war games, and as a tool to compare performance of competing sensor systems for a specific task. A comprehensive TA model is the Target Acquisition Model (TARGAC), developed at the U.S. Army Research Laboratory, Battlefield Environment Directorate (ARL-BED). The model is part of the Electro-Optical Systems Atmospheric Effects Library (EOSAEL).

Target recognition and identification by human observers is a complex pattern recognition process, that is not yet fully understood. Models that describe this human capability must still rely on a number of simplifying assumptions. In addition to the human pattern recognition capabilities, psychological factors play an important role in human performance, and these factors are even more difficult to incorporate into a model. Therefore, the quality of the models is not always known. To allow correct and valid use of TA models, it is necessary to know how well the models predict TA performance and under which conditions a model may be used. Furthermore, it is important to know the confidence limits of the predicted model output. Also, the fact that the accuracy of the model predictions is usually not known gives rise to the so-called false precision problem, which means that values for which the accuracy is not known are treated as being exactly correct. All this means that a careful validation of a TA model should be carried out before conclusions can be drawn from its predictions.

Model validation can be done by analyzing the structure of the model, or comparing the predictions of the model to the actual performance of observers

in military (field) tasks. This study takes the latter approach, by measuring observer performance on a large number of thermal images collected during the field trial Battlefield Emissive Sources Trials under the European Theater Weather and Obscurants (BEST TWO) (organized by the North Atlantic Treaty Organization (NATO) AC243/Panel4/RSG.15, 1990) and calculating the corresponding TARGAC predictions.

This study was carried out by the Perception Department of The Netherlands Organization for Applied Scientific Research (TNO), TNO Human Factors Research Institute (TNO-HFRI) in Soesterberg, The Netherlands. TNO is a large independent research facility serving the Dutch government and the Dutch Ministry of Defense.

1.1 Observer Experiments and Field Trials

Observer experiments must be carried out, using carefully controlled procedures, according to a design that allows proper statistical analyses of the data. For this reason, such experiments are best carried out in the laboratory where conditions can be controlled and repetition of identical experiments with different observers is possible. In a field situation, conditions are difficult to control and often change quickly. Further, it is difficult to obtain accurate and reliable data on observer performance.

BEST TWO provided an opportunity to collect imagery for laboratory observer performance experiments that would yield data with sufficient accuracy for a quantitative evaluation of TA models. During the test, a large amount of thermal images of stationary and moving target vehicles at many distances were recorded. Target recognition and identification performance were determined for these images. A limited observer experiment was carried out in the field for validation of the observer scores measured in the laboratory. Also collected during the trial were meteorological data, target contrast values, and other parameters required by TARGAC to make acquisition range predictions.

1.2 Overview

The work carried out within this project consists of three studies. The experimental methods are described in sections 2 and 3, and the results are discussed in sections 4, 5, and 6. These sections are based on a number of

mostly confidential reports published earlier (see bibliography). Most of the data in the original reports are confidential because they reveal recognition and identification performance for thermal imaging of a number of actual military targets. To make the present report unclassified, the targets have been coded with the letters A through I in all instances concerning data and graphs. This does not in any way influence the conclusions that can be drawn from this study. The key to the target codes is available on request from TNO-HFRI or ARL-BED (Dr. P. Gillespie). In addition to coding the target vehicle names, the contents of the original reports have been slightly rearranged to obtain a coherent final report.

1.2.1 Field Test

In section 2, an overview of the NATO BEST TWO is presented, and the scenarios carried out in the field and the different recording conditions are described. The weather during the test was hot and dry and very constant over the 3-week test period, meaning the effect of the weather on TA performance could not really be studied.

1.2.2 Design of Laboratory Experiments

Section 3 treats the design of the laboratory observer experiments and the training and selection of the observers. It is shown as follows:

- For the restricted set of target vehicles used, observers without experience could be trained to an acceptable level in a few hours.

- Large differences in performance between subjects occurred, and within a few hours good observers could be distinguished from poor ones. A few observers were exceptionally good, and about 30 percent of the observers were unsuitable for the task.

- Military and civilian observers were tested and no overall difference in performance between the two groups was found. Section 5, however, shows a difference in response behavior.

1.2.3 Study I: Observer TA Performance

Section 4 presents the results of Study I. Target recognition and identification probability was determined as a function of range, for a number of different conditions. The effect of target type, time of day, approach route, and target motion on the observer scores was determined. Further, two different ways of presenting the targets were used: pop-up and sequential. In the pop-up presentation mode, randomly picked targets appeared at random positions in the field; in the sequential mode, the targets were presented as an ordered sequence of decreasing distance, from 4 km down. The latter condition simulated a target approach, during which the observer may accumulate information on the target. Both ways of target presentation have military significance. Search and target detection were not studied.

The data were collected in two main experiments using 24 observers. A total of 811 different images containing single targets were presented, and each target presentation was repeated five times at random in the course of an experiment. All datapoints in the graphs are the averaged scores over all participating observers. Appendix A presents a complete overview of all the observer performance data collected.

The results can be summarized as follows:

- Identification and recognition performance varies considerably for the different targets, in different trials, under different conditions; in some cases targets were recognized at 4 km, while in others they could not be recognized at 1 km.

- Identification and recognition performance for a target that suddenly appears at a certain location (pop-up) may be considerably worse than for a target approaching from a large distance (sequential). Current TA models do not take this difference into account.

- Head-on target motion, with or without a dark dust cloud behind the target, does not have a large influence on identification or recognition performance. However, it may have an effect on detection probability.

1.2.4 Study II: Reliability of Observer Responses

Section 5 describes Study II, in which the reliability of observer responses and some of the more psychological and task-structure effects on the observer performance are analyzed. In a TA task, it is always possible for an observer to make a wrong judgement. This can make the observer feel unsure and hesitant to give a report, although the target is correctly recognized. On the other hand, if the observer is confident enough to give a report or to undertake an action, it is of obvious importance to know the probability of being wrong and the factors influencing this probability. Therefore, apart from the observer's skill in identifying or recognizing a target, observer confidence and the corresponding response behavior forms an important factor in TA performance. This study was designed so the influence of skill and behavior on TA performance could be separately analyzed. Also, the reliability of first reports (the first time an observer reports an identification or recognition) during a target approach was analyzed. The results show the following:

- Observers possessing the same skill can differ widely in behavior. In practice, this means that one observer gives reports at much closer target ranges than another, although, in principle, the observers could provide the same information at the same distance.

- There is a large range of target distances at which the observers do not feel sure enough to report an identification, but respond correctly if they are forced to. Thus, if the circumstances ask for it, acquisition range may be significantly increased by forcing the observers to respond. Because this may also lead to an increase of false alarms, the instructions should be given depending on the circumstances.

- The reliability (the probability of being correct) of a first identification or recognition report during a target approach, is 80 percent, averaged over all subjects. This percentage is largely independent of target distance, target type, part of day (POD), and target background. However, large individual differences (between 55 and 97 percent) are found.

1.2.5 Study III: TARGAC Validation

Section 6 presents the results of Study III. A comparison is made between TARGAC recognition predictions and measured observer performance for a large number of trials.

First, TARGAC was subjected to sensitivity analyses in which the effect of changes in the input parameters on the model output were determined. This showed that, for BEST TWO, the predicted recognition ranges are almost independent of time of day and target and background temperatures. Only the target effective dimension and the thermal imager resolution limit (the cut-off frequency of the minimum resolvable temperature difference curve (MRTD)) had a significant effect on model output. The result was, in fact, caused by the excellent atmospheric conditions during BEST TWO. It means that, in this study, the recognition performance predictions are determined solely by the system performance module of TARGAC, which is equivalent to the 1-dimensional Night Vision Electro-Optics Sensors Directorate (NVESD) Static Performance Model.

The model was validated by determining the ratio between measured and predicted acquisition ranges for a large number of trials (a large number of target approaches in different conditions). The results are expressed as a probability distribution of this ratio. The mean of this distribution quantitatively shows how well the model predicts overall acquisition performance over a large number of trials. The variance of the distribution is a quantitative measure of the accuracy of the model predictions for individual trials. If the mean of the distribution is equal to 1 and the variance is equal to 0, the model predictions are perfect. Deviations of the ideal values indicate the extent to which to the model can be used. The results for TARGAC BEST TWO are as follows:

The mean of the ratio distribution is 1.8 (+/- 0.2), which means that, on average, observer performance is considerably better than the model predicts, and a correction factor of 1.8 should be used to match the predicted recognition ranges to the results of the observer experiments. However, the shape of the mean probability versus range curve is very similar for the model and the observations.

The variance of the ratio distribution is large: the ratio between measured and predicted acquisition range spreads between 0.9 and 3.6 (95 percent level) (spans a range of a factor of 4). Thus, TARGAC predictions of recognition range for individual targets can differ from the actual recognition range by a factor between 0.9 and 3.6, which means that TARGAC does not predict recognition performance very well. A similar conclusion may hold for other models based on the same principle.

The TARGAC output was compared to the 1-dimensional (1-D) ACQUIRE target acquisition model developed by NVESD, which showed that both models produce identical recognition versus ranges curves. Both models underpredict the mean recognition range by a factor of 1.8. Predictions of the newer 2-dimensional (2-D) ACQUIRE model, which uses the Johnson criteria in 2 dimensions, yielded a much better result: the factor of 1.8 was reduced to about 1.3. However, the variance in ratio distribution remained the same when the 2-D model was used.

The sensitivity analysis and a number of tests performed on TARGAC before the actual validation was carried out, brought to light a number of problems and errors in the software. The errors were fixed after consultation with Dr. Patti Gillespie from ARL-BED, before the validation was carried out.

1.3 Conclusions and Recommendations

TARGAC was validated using BEST TWO observer performance data for recognition of targets in front view. The major conclusions of Studies I and II are that human acquisition performance depends considerably on factors such as target structure, local terrain structure, and cognitive factors.

In TARGAC, and in many other TA models, these factors are not incorporated. Study III shows that the TARGAC predictions for BEST TWO hardly depend on the experimental and field conditions. Therefore, important differences between measured and predicted recognition performance were found.

The main results of the validation follow:

1. The model predictions are too conservative. On average, TARGAC underestimates recognition range by a factor 1.8 (+/- 0.2).

2. TARGAC makes accurate predictions for individual cases. The analyses show that the uncertainty interval roughly ranges from 0.9 to 3.6 times the predicted acquisition range, spanning a range of a factor 4.

3. TARGAC recognition performance predictions for BEST TWO (excellent weather) are determined solely by its system performance module, which is equivalent to the 1-D NVESD Static Performance Model.

4. The TARGAC predictions for overall mean performance can be improved by incorporating the 2-D version of the Static Performance Model. For individual cases, however, the predictions with the 2-D version are not better than those with the 1-D version.

5. It is proposed that TARGAC predictions are not only presented as single numbers for acquisition probability versus target range, but that some indication is given of the accuracy of the results, preferably in the form of a 95 percent confidence interval.

6. The version of TARGAC tested (PC version, released in June 1992) contained a number of software errors and some minor problems. A number of corrections are suggested. Additional work in modularizing and streamlining the model is recommended. It is also recommended that the model is given a more consistent and user-friendly interface.

7. TARGAC and other models that incorporate the NVESD Static Performance Model should only be used to provide an indication of the actual acquisition performance.

2. BEST TWO

2.0 Summary

BEST TWO was held in Mourmelon, France in Jul and Aug 90. The test was organized by NATO AC/243, Panel 4, RSG 15; the participating countries were the United States, England, Germany, France, Denmark, and The Netherlands. The purpose of BEST TWO was to quantify the performance of EO imaging and observation devices under battlefield conditions. This section gives an overview of the four scenarios carried out.

In Scenario 1, single target vehicles (tanks, armed personnel carriers, and trucks) drove down a predefined track from a distance of 4000 m toward the main instrumentation area (MIA). The targets stopped for 2 min at designated positions along the track (roughly every 300 m). Two different tracks were used: one with 16 stop positions and one with 10 stop positions. This scenario allowed the recording of imagery of stationary targets at a range of distances between 4000 and 1000 m.

Scenario 2 was very similar to Scenario 1, the difference being that the target vehicles did not stop. Battlefield effects were included in some versions of Scenario 2.

In Scenario 3, a column of 8 or 12 target vehicles drove along a track across the field of view. The purpose of Scenario 3 was to study the recognition of groups of targets and the effects of dust on observer/system capability.

In Scenario 4, an attack formation of four tanks and eight armored personnel carriers (APCs) approached the MIA from a distance of 4 km. The purpose of Scenario 4 was to assess the effects of simulated artillery barrages and/or smoke on observer/system performance in a realistic task environment.

2.1 Introduction

BEST TWO was held at Camp de Mourmelon, near Chalons-sur-Marne, in the Northeastern part of France, from 26 Jul to 10 Aug 90. This report describes the four scenarios carried out during BEST TWO. A detailed description of target vehicle order, timing and movement during the scenarios, and the timing of the simulated battle events is presented elsewhere. [1]

2.1.1 General Information

2.1.1.1 *Session organization.*— BEST TWO was organized in 4-h sessions, and two or three sessions were carried out per 24-h day. The session time slots are listed in table 1. The sessions were numbered by a code xx.y where xx is the day of the month and y is the slot number (table 1). Because the experiments were carried out in the last week of July and the first 10 days of August, use of the day numbers only allows a unique session code (session 27.3 is the afternoon session on 27 Jul and 3.4 is the late night session on 3 Aug).

Table 1. Session time slots

Slot number	Time period
1	0200 - 0600 (early night)
2	0900 - 1300 (morning)
3	1400 - 1800 (afternoon)
4	2200 - 0200 (late night)

2.1.1.2 *Terrain layout.*— The basic pattern of the experiments was that at least one target vehicle was stationary or moving in the terrain while measurements and recordings were made from several sites in the field. Seen from the MIA, the terrain was roughly 2-km wide and 4-km long. In most scenarios, the target vehicles approached the MIA from the far end of the field up to 1 or 2 km from the MIA. Maps showing the target vehicle routes are presented for each scenario in section 2. The maps were made with the field survey data provided by France. The target vehicle routes were marked in the field with numbered signs. The numbers made it possible to repeatedly position the targets at the

same designated points, as in Scenario 1. The signs were dimly illuminated at night to allow the target vehicle drivers to find their way. Different colored signs were used for the routes for the different scenarios: white and black for Scenario 1 and 2 (left and right), yellow for Scenario 3, and blue, blue/black, and black/blue for Scenario 4. Sometimes the approach routes are named after the color of the signs used.

2.1.1.3 *Test time.*— Standard test time was provided by France and distributed in IRIG-B format to all national setups in the MIA on a coax cable. This signal was recorded with the measurements to provide a common time base for all data. The code was recorded on the audio channel of the video systems used to record imagery.

2.1.1.4 *Laser experiments.*— Lasers were used during a large number of experiments, and during about half the sessions all personnel in the field were required to wear laser safety goggles.

2.1.1.5 *Characterization.*— In addition to the four scenarios described in this report, there were three characterization sessions in which physical measurements of isolated battle effects were made. One or two tanks were used as targets during these measurements, but no large scale vehicle movements were involved.

2.1.2 Target Vehicles

A total of 14 target vehicles of three different types were used: two types of tank, three types of APC, and one type of truck. The French supplied 3 AMX-30 tanks, 6 AMX-10 APCs and 2 trucks. The Dutch contributed one Leopard 2 tank and two versions of the YPR 765 APC: the YPR-PRI, a standard APC with a 25-mm cannon, and the YPR-PRAT that has a dualtube launched, optically tracked, wire-guided missile (TOW) anti-tank missile turret. Germany added special thermal camouflage materials to three of the French vehicles; these are identified with a C behind their name. Table 2 lists the target vehicles used.

Table 2. Target vehicles

Tanks	APCs	Wheeled
Leopard 2	PRI	Truck
AMX-30	PRAT	Truck C
AMX-30 C	AMX-10	
	AMX-10 C	

To have an operational signature during the test, the drivers exercised their vehicles for 15 to 30 min before each run outside the view from the instrumentation areas. There was a signature post at the back of the field where the Danish and the Germans took calibrated measurements of all target vehicles before each run. If the signature was not right, the driver had to go back to further warm up the vehicle.

Target vehicle management was done by a cadet of the Dutch Royal Military Academy, who was in radio contact with the general test management in the MIA.

2.1.3 Battle Events

During the test, a large number of simulated battlefield effects were produced:

a) sandbags - a simulation of an artillery barrage by exploding sandbags that were suspended from tripods, each sandbag representing two 122-mm and one 152-mm Soviet shells

b) LUST - (Limited use smoke technology device) canisters producing white phosphorous smoke

c) fires - a simulation of a burning object on the battlefield by burning fuel and tires in oil drums

Details of the execution of the simulated battle events are described in Danielian. [2] Scenarios 2, 3, and 4 were carried out a number of times, with and without several of the simulated battle events.

2.2 Scenarios

2.2.1 Scenario 1

In Scenario 1, single target vehicles drove down one of two different approach routes (left and right). The routes are shown on the map in figure 1, which also shows the location of the MIA. The numbers along the tracks are the locations of the numbered signs. The distances between the stop positions and the MIA are given in Valeton, Bijl, and De Reus. [1] In this scenario, the target vehicles stopped for about 2 min at each stop sign, allowing the teams in the instrumentation areas to make recordings and measurements of stationary targets. A stationary target in this case is not only a target that does not move, but also a target that has no large dust cloud behind it. The left route came as close as 1000 m from the MIA, while the closest distance for the right route was about 1600 m. Note that stop sign 14 is missing on the left route. It disappeared during the first day of the test. Scenario 1 was carried out six times: three times along the left route and three times along the right route. No battlefield effects were included.

Several nations had observers in the MIA doing a real-time TA task. The order of the targets in each session was chosen at random by the test management so the observers never knew in advance which target they were going to see. A session lasted 4 h, each target vehicle run lasted about 30 min, and on average 6 to 8 targets could participate in each session.

2.2.2 Scenario 2

Scenario 2 was very similar to Scenario 1, the only difference being that the target vehicles drove down the approach routes continuously, without stopping. The location of each vehicle during a run was later determined (appendix A). This scenario allowed recordings and measurements of moving targets. Scenario 2 was carried out six times under different conditions. Two target vehicle speeds were used: 20 km/h (fast) and 8 km/h (slow). The purpose of the two speeds was to create one condition of moving targets with dust thrown

up by the vehicles, and one condition of moving targets without dust thrown up by the vehicles.

In practice, it appeared that there was little difference in the amount of dust produced by the targets in both conditions. Two kinds of battlefield effects were used: fires and artillery dust simulations. Table 3 gives the conditions used for the six versions of Scenario 2.

In Scenario 2C, one sandbag was exploded 50 m to the left of the route, and one sandbag was exploded 50 m to the right of the route. Both sandbags were exploded at 3500 m and at 2500 m during each target vehicle run. The timing was such that the explosions went off when the target was right between the two bags.

In this scenario, the targets could complete a run in less, sometimes much less, than 30 min. However, to coordinate target movements with battle events and the timing of helicopter and fixed wing aircraft overflights, the targets were scheduled to start at exactly the half hours on the clock. This meant that usually only seven runs could be completed; therefore, not all targets could be included in each session.

Table 3. Details of Scenario 2

Scenario	Route/Speed	Session No.	Battle Effects
2A	left/slow	31.4	none
2A	left/fast	31.3	none
2A	right/slow	8.1	none
2B	left/slow	3.4	6 fires
2B	left/fast	1.3	6 fires
2C	left/slow	31.2	2 sandbags at 3500 m 2 sandbags at 2500 m

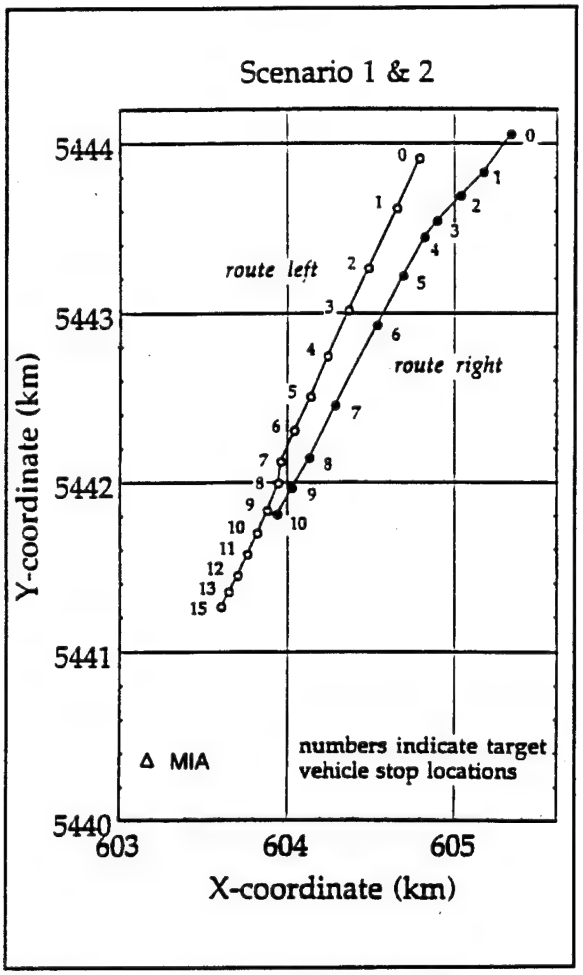


Figure 1. Approach routes for Scenarios 1 and 2.

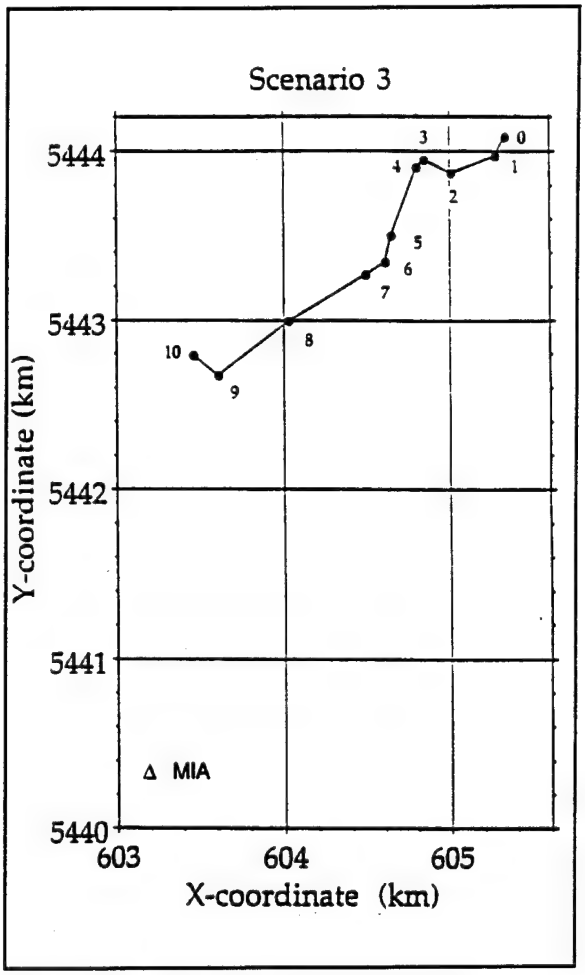


Figure 2. Approach route for Scenario 3.

2.2.3 Scenario 3

In Scenario 3, a column of target vehicles at a mutual distance of 50 m drove down an oblique track across the field, see the map in figure 2. The vehicle speed was 20 km/h.

The size and composition of a column of vehicles is important information. When a column is partly obscured by dust or battle effects, correct appraisal of the situation might be hampered. The purpose of this scenario was to determine whether observers can classify a column on the basis of its whole appearance, or gestalt, instead of by seeing all individual vehicles.

To test this idea, four main types of columns were used. All columns consisted of tanks and APCs. The tanks were always grouped together. The columns were short (8 targets) or long (12 targets). In the short columns, the tank group was in front or at the back; in the long columns, it was in front or in the middle. The four formations are illustrated in figure 3. The short column was typical of two platoons moving to contact; the long column was typical of a company-size column of mixed composition. The distribution of the available tanks, APCs and trucks within the layout of a given column was different for each run.

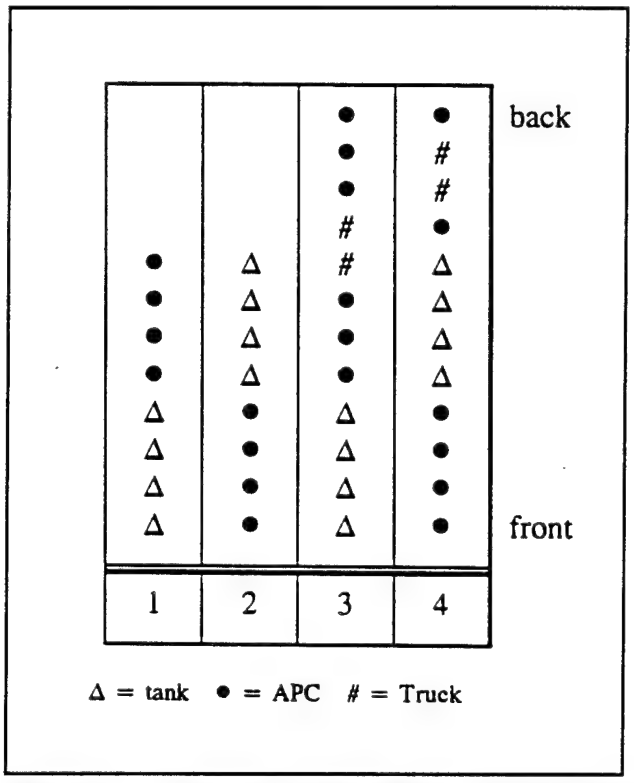


Figure 3. General composition of the four column types of Scenario 3. The layout is presented as the targets were seen coming toward the MIA; the bottom ones are in front.

Scenario 3 was carried out three times, twice without battlefield effects and once with three fires located at 3500 m from the MIA along the track. Details are presented in table 4. In each session, four runs could be completed, so a total of 12 runs were realized during the whole test.

A number of observers from different nations made real-time observations for Scenario 3. The task of the observers was to first quickly determine whether a column was long or short, second classify the type (tanks in front or not), and third name the targets in the column.

Table 4. Details of Scenario 3

Scenario	Session No.	Battle Effects
3A	7.3	none
3A	8.2	none
3B	6.3	3 fires at 3500 m

The main conclusion from the field observations is that the columns in this test are never recognized at a glance; the idea that a column with a certain mission has a certain gestalt does not appear to be true. The observers judge each vehicle of a column as it comes into view, and they count them one by one to find out what kind of column they are dealing with. The columns throw up a lot of dust, and depending on the wind, a whole column may be obscured by the dust cloud generated by the first vehicle.

2.2.4 Scenario 4

2.2.4.1 *General.*— In Scenario 4, an attack formation consisting of four tanks followed by eight APCs approached the MIA from the far end of the field to about 2500 m. The speed of the vehicles was 15 km/h. The purpose of this scenario was threefold: (1) to collect data on the number and duration of holes or visibility windows in dust and/or smoke clouds; (2) to determine how many targets of an attack formation can be seen at any one time; (These two points are important for determining the effectiveness of guided missiles like the TOW and direct fire systems like tank and APC guns.) and (3) to record realistic

thermal and visual imagery of an attacking formation under various battlefield circumstances for training purposes. This scenario was carried out 10 times in five forms, and usually several runs could be made in one session. The different versions and the combinations of battlefield effects used are listed in table 5.

Table 5. Details of Scenario 4

Scenario	Session No.	Battle Effects
4A	27.2-30.3	None
4B	6.2	Rolling Artillery Barrage near targets, see text
4C	1.2-7.2-10.3	Artillery barrage in front of MIA
4D	2.3-9.2	WP smoke in front of MIA
4E	3.2-9.3	Artillery barrage + WP smoke in front of MIA
4X	9.3	None

Scenario 4A was a baseline condition for Scenarios 4C, D, and E. Scenario 4B was different from the others and is described separately. The target tracks and explosion areas are presented in figures 4a and 4b. Scenario 4X was a short search experiment and was different from all other versions of scenario 4.

2.2.4.2 *Scenario 4 A, C, D, and E.*— In Scenarios 4A, C, D and E, the attack formation was about 600-m wide. Four tanks were driving down the four approach routes shown in figure 4a, and two APCs followed each tank at a distance of 100 to 200 m. One of each pair of APCs drove 50 m to the left of the tank-track, the other drove 50 m to the right. The formation is schematically depicted in figure 5.

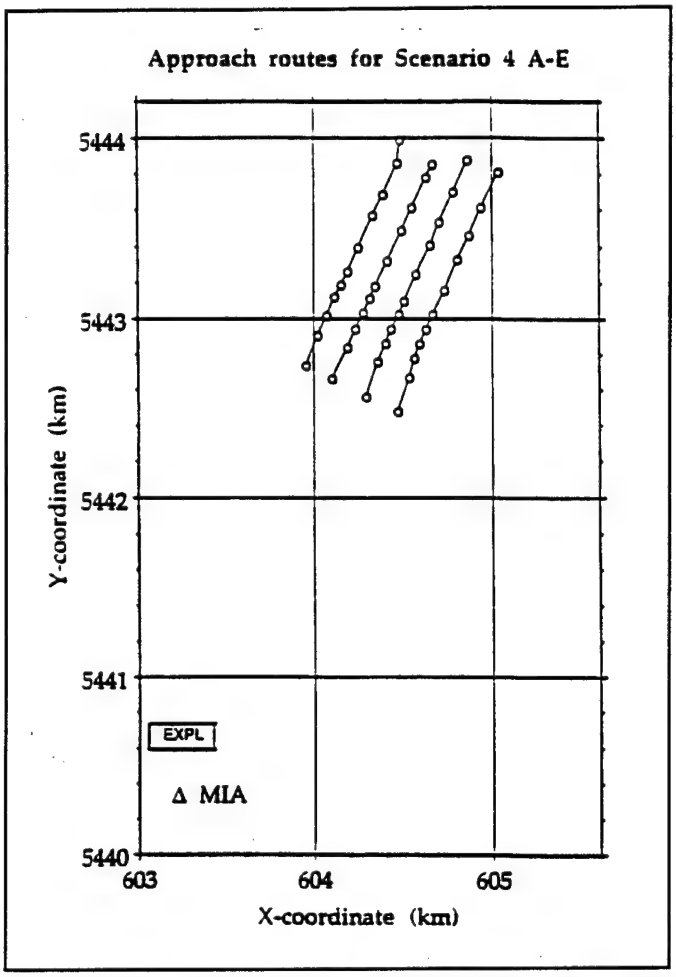


Figure 4a. Approach routes for scenarios 4A, C, D, and E. The explosion area is marked EXPL. Mutual distance between the tracks is 200 m.

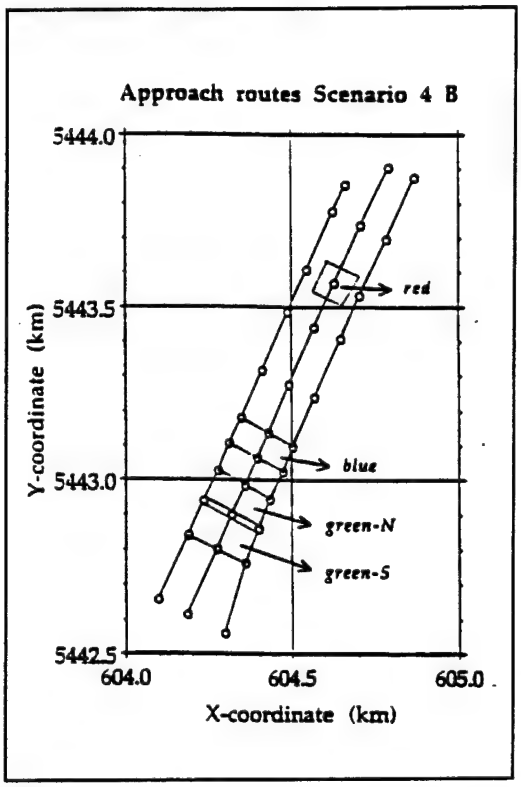


Figure 4b. Approach routes for scenario 4B. The four explosion areas are indicated as red, blue, green-N, and green-S. The scale is twice that of figure 4a.

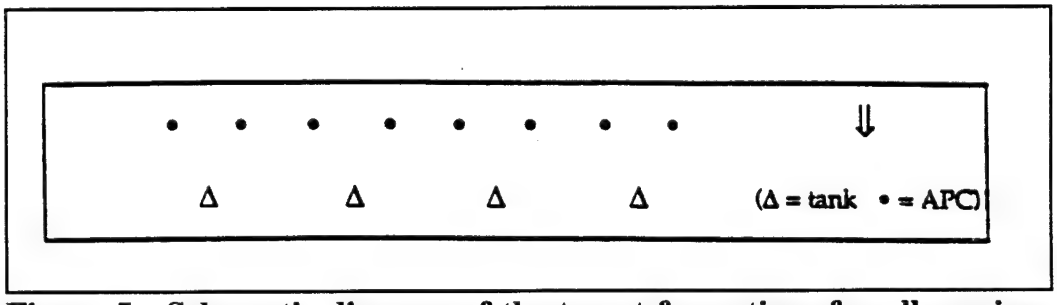


Figure 5. Schematic diagram of the target formations for all versions of Scenario 4. The arrow on the right indicates the direction of motion.

The battle effects were produced in a 400- by 100-m explosion area (marked EXPL on the map in figure 4a) at a distance of 300 m from front of the MIA.

The scenarios typically lasted 6 to 8 min. To illustrate the timing of target movements and battle events, a sample of a planning sheet used for these scenarios is presented in figure 6. The scenario began after all targets were warmed up and in position at the beginning of the tracks. During the first 3 min, battle events (sandbag explosions, LUST devices, or both) were set off at 1-min intervals in the explosion area in front of the MIA; the vehicles waited. After the dust/smoke screen was built up, the formation started to move. Battle events were generated once a minute, for 6 min, to sustain the obscuration. By the time the tanks reached the end of their tracks, at about 2500 m from the MIA, the event was over.

2.2.4.3 *Scenario 4B.*— A rolling artillery barrage was simulated in Scenario 4B. The idea is that artillery paves the way for an attack formation by shelling from behind the area in front of the troops to neutralize all enemy defenses to enable fast and unobstructed advance of the forces.

The attack formation was the same as in figure 5. In the first phase (red), the formation was 200-m wide; after that it expanded to 600-m wide. Figure 4b shows the tracks and explosion areas. The figure shows an enlarged (2x) view of the target area. The two outer tracks in figure 4b are the same as the two inner tracks of figure 4a. The central track in figure 4b was added for Scenario 4B and coincided with the left route for Scenarios 1 and 2. The four tanks were evenly spaced over the 200 m width: one tank followed the leftmost track, one followed the rightmost track, one drove 30 m to the right of the central track, and the other drove 30 m to the left of the central track.

Four explosion areas were laid out in the field, and they were labelled red, blue, green-N (north), and green-S (south). In the four explosion areas, different area/time densities of artillery bombardment were simulated (table 6). In area red, 20 sandbags, spread evenly over the 100 by 100 m, were exploded over a period of 2 min, while in both green areas the same number of bags was spread out over twice the area and exploded in a quarter of the time. These different conditions threw up dust clouds of different densities and duration.

The execution of this scenario is described below. The position of the formation is indicated as 3700/3900 m, where the numbers represent the distances of the tanks/APCs from the MIA.

Exp No. 1.2		Scen: 4C				
Rel Time (min)	Abs Time (h:min)	EXPLOSION CREW Sandbags			TARGET FORMATION (Vehicles use Blue/Black # signs)	
-50						Warm up of all vehicles
-10						Thermal signatures of 1 Tank & 1 APC
-5		Pre-check charges				Tanks to sign # 1; APC's to sign # 0
-4						
-3				Tank	APC	
-2				Pos. - Dist.	Pos. - Dist.	
-1						
0		Explode 1-st volley	(21 bags)	1 - 3700 m	0 - 3900 m	All vehicles WAIT
1		Explode 2-nd volley	(7 bags)	1 - 3700 m	0 - 3900 m	All vehicles WAIT
2		Explode 3-rd volley	(7 bags)	1 - 3700 m	0 - 3900 m	All vehicles WAIT
3		Explode 4-th volley	(7 bags)	1 - 3700 m	0 - 3900 m	All vehicles START; 15 km/h
4		Explode 5-th volley	(7 bags)	3 - 3400 m	1 - 3700 m	
5		Explode 6-th volley	(7 bags)	4 - 3150 m	3 - 3400 m	
6		Explode 7-th volley	(7 bags)	7 - 2850 m	5 - 3100 m	
7		Explode 8-th volley	(7 bags)	9 - 2650 m	7 - 2850 m	
8		Explode 9-th volley	(7 bags)	10 - 2400 m	9 - 2600 m	All vehicles WAIT
9						
10						
11						
12						All vehicles take tactical front-covered
13						positions facing MIA. Wait for 5 min.
14						
15						
16						
17		*** END SCENARIO ***				All vehicles leave field in sir.
18						followibg route Yellow.

Figure 6. An example of a planning sheet: Scenario 4.

Table 6. Artillery barrage density in Scenario 4B

Expl. Area	Size	No. Sandbags	Duration
red	100 x 100 m	20	2 min
blue	200 x 100 m	20	1 min
green-N	200 x 100 m	20	0.5 min
green-S	200 x 100 m	20	0.5 min

2.2.4.4 *Event RED.—*

1. The target formation was stationary in position at 3700/3900 m.

2. The explosions in area red were set off; the targets did not move.

2.2.4.5 *Event BLUE.—*

1. The formation was moved 200 m down the tracks, to the blue start position at 3500/3700 m.

2. The targets started driving down the tracks at 15 km/h (250 m/min).

3. As soon as the formation was in motion, the explosions were started in area blue.

4. During the explosions the formation drove about 300 m; the formation stopped when the explosions were finished in area blue.

2.2.4.6 *Event GREEN.—*

1. The formation was moved 300 m down the tracks to the green start position at 3200/3400 m.

2. The targets started driving down the tracks at 15 km/h (250 m/min).

3. When the formation started moving, the explosions were started in area green: the first 1/2 min in area green-N, the next 1/2 min in area green-S.

4. During the explosions the formation drove about 250 m; when the targets reached the end-point at 2900/3000 m the blue explosions were finished in area green.

Analyses of Scenario 4 are presented in Vonhof and Goessen. [3]

2.2.4.7 *Scenario 4X.*— In this version of Scenario 4, the attack formation took position along the ridge with some trees and bushes about 2000 m from the MIA. The vehicles were to take a position suitable for an attack on the MIA but obscured from view by using the local terrain features as much as possible. The task of the observers in the MIA was to find the targets on the thermal imagers. This scenario was carried out only once, to get an idea of the detection probabilities in the described situation.

The surprising result of this short test was that only about half the targets were found, and half of the responses given turned out to be false alarms. [4]

2.3 Target Positions in Scenario 2

For the analyses of the data collected during the experiments, the location and distance of each target must be known at all times. Because obtaining this time/distance information for Scenario 2 was not straightforward, the procedure used is outlined below.

2.3.1 Radar Data

The U.S. fielded a multitarget tracking system (AUTOFEDS) that was developed by NVESD for recording target vehicle movements and observer responses during field trials to provide all target vehicle locations in x,y coordinates relative to the MIA as a function of time. Unfortunately this system was not operational during most of the test, and the limited data available proved unusable. Fortunately, two French teams had a battlefield radar: 1) LMT-Radio Professionelle (French battlefield radar system) from Boulonge and 2) RASIT (French battlefield radar system) from ETCA (Etablisement Centrale Technique d'Armement), Arcueil. The radar systems were to provide backup information on target vehicle locations; therefore,

information, albeit only for moving single targets (Scenario 2), was available after all.

Two problems emerged in the analyses of the data provided by both radar systems. First, it appeared that in most of the data of the LMT radar, the compass bearing was not set according to the local deviation of the earth's magnetic field. By rotating the coordinates of the vehicles by the amount of the local deviation, a reasonable fit with the land survey data could be obtained. This exercise proved that the LMT data were, in principle, usable. The RASIT radar produced geometrically correct data.

A second problem in the radar data concerns the time information. The master time for the test was provided by France and made available to all nations in the MIA in IRIG-B format on a coax cable. The IRIG-B signal was recorded on the audio channel of video recorders; therefore, the correct time for the imagery on the video. The computers of the radar systems had no hard-wired connection to the IRIG-B signal, and the computer clocks were presumably set by hand. This resulted, for many sessions, in differences between the time axes of the LMT radar, the RASIT radar, and the IRIG-B time. The problems were noticed when time/distance information of the LMT and RASIT radars for a few runs were compared with actual target time/position as inferred from the video image/audio channel. Time differences between IRIG-B and the two radar systems of up to 2 min have been found. Because the radar data was not considered reliable, the time/position relation of all targets in Scenario 2 were determined by direct analyses of the video imagery.

2.3.2 *Visual Determination of Target Time/Distance Profile*

The images of a complete Scenario 1 run were displayed on a monitor. A transparent sheet was attached to the face of the monitor, the approach route was sketched on the sheet and the locations of stationary target vehicles were marked. The marks correspond with the locations of the numbered signs along the approach route. This procedure was done for the left and right approach routes. Next, imagery of Scenario 2 runs was played back on the same monitor while the IRIG-B time was displayed. At each instance when a target, as it moved along the track, passed a mark on the monitor face, the target was

at a stop-sign location. The IRIG-B time of the instance was noted, and because the distance of all numbered signs along the track are known, a series of correct time/distance values for each run was obtained. These data are indicated as visual or VIS. The VIS time/distance data were plotted, obvious errors were corrected, and comparisons with the radar data were made. A complete set of time/distance plots is presented in Valeton, Bijl, and De Reus. [1]

As an example of the differences in time/distance relations found for the three data sets, two extremes are presented in figure 7. In figure 7a, a run is plotted for which the time/distance relations from all three sources coincide perfectly. Figure 7b shows an example in which there are time shifts of 1.5 and 2 min between IRIG-B and the radar data. Each symbol on a radar curve indicates a position generated by the radar. Note that the curves are almost straight lines, showing the target vehicle drove at constant speed and the three data sets are consistent with the only difference being a shift in the time setting.

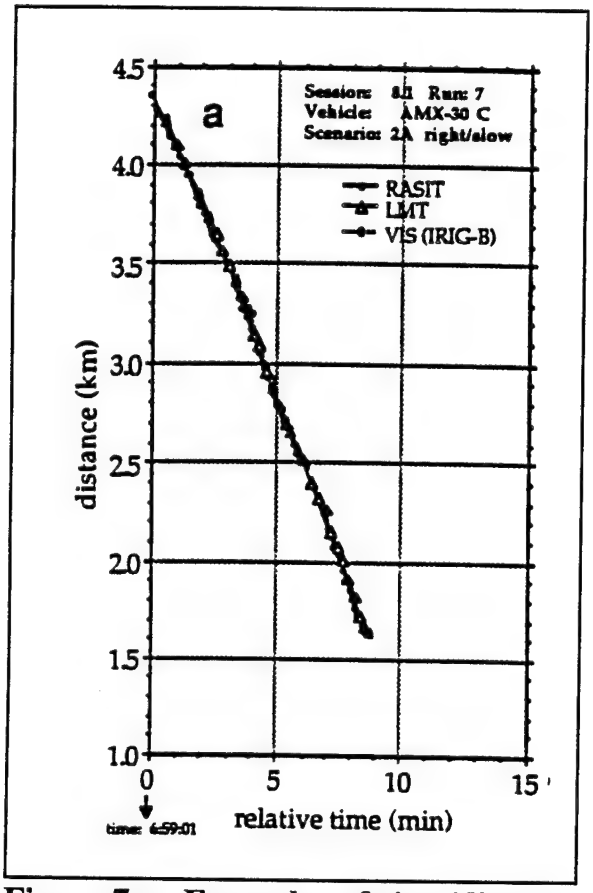


Figure 7a. Examples of time/distance relations for Scenario 2 from three different sources: the RASIT radar, the LMT radar, and visual data obtained from IR imagery. The data from the three sources coincide.

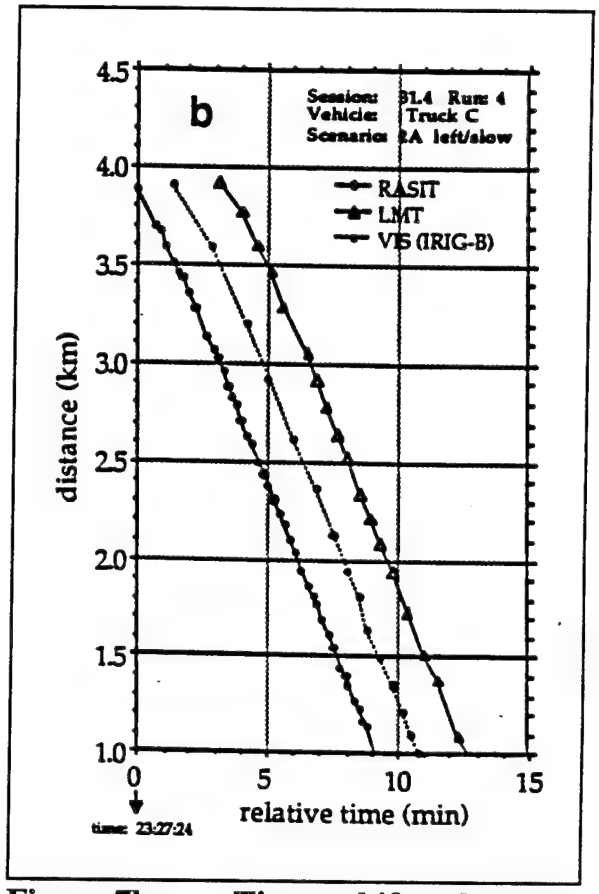


Figure 7b. Time shifts between IRIG-B time and both radar data sets are apparent. Time shifts indicate errors in the time setting of the radar systems.

2.4 Closing Remarks

BEST TWO was executed with great success. Only one of the planned sessions had to be canceled, because the schedule was overloaded. All national teams expended a great effort and collaborated successfully. In addition, France supplied field radio systems, the standard test time, a safety officer, road blocks, and the explosives team that prepared, set up, and executed all battle event simulations with a very high success rate. Further, a large number of logistic problems were solved every day. The weather during the test was

mostly constant; very hot and dry. Meteorological data are presented in Smith and Corbin. [5]

In hindsight, it might be concluded that the test schedule had one flaw. Night sessions were to be included in BEST TWO, but it was decided not to sacrifice any daytime experiments.

3. Design of the Observer Experiments

3.0 Summary

During BEST TWO, images of single stationary targets (Scenario 1) and moving targets (Scenario 2) were recorded with a thermal imager. The images are used in observer performance experiments to collect data for the evaluation of TA models and for operational purposes. This section describes the methods and design the experiments.

Stimuli consisting of short image sequences were shown to observers, using an analogue video disc system. This system allows presentation of events in random order at fast pace, while retaining the dynamic character of the imagery, for both stationary and moving targets. Identification and recognition performance was determined as a function of target range.

An extensive training program was developed; transfer of training to the main experiments was satisfactory. Large differences in performance between subjects occurred. The criteria for selecting observers for further analyses are discussed. Military and civilian observers were tested, and no overall difference in performance between the two groups was found.

3.1 Introduction

Thermal images recorded during BEST TWO were used in observer experiments in the laboratory to determine TA performance. The purpose of these experiments is to collect data for the evaluation and development of TA models and supply rules of thumb for operational purposes. The present observer performance experiments were restricted to target identification and recognition; target detection and search are not studied. No battlefield effects were included.

This section treats the design of the experiments, the setup that was built, the observer training, and the subsequent observer selection process. Also, a

comparison is made between the performance of military and civilian observers.

Because detailed information on thermal image recognition on target vehicles is confidential, target names are coded with the letters A through I in all data plots. This does not in any way affect the conclusions that can be drawn from the experiments (see section 3.5.3). The key to the vehicle code is available on request.

3.2 Methods

3.2.1 Field Recordings

Four scenarios were carried out during BEST TWO. An overview of these scenarios has been presented in section 2. Detailed information on the events during the trials including test schedules, maps, time tables, vehicle positions, and battlefield events is reported in Valeton, Bijl, and De Reus. [1] The laboratory experiments were conducted with imagery collected during Scenarios 1 and 2. Recordings were made of stationary and moving single target vehicles at a range of distances between 4000 and 1000 m. Nine target vehicles were used in these scenarios: three tanks, four APCs, and two wheeled vehicles. Three vehicles were camouflaged.

The imagery was recorded from an 8 to 12 μm thermal imager on Umatic video tape. The field of view was 5 $\times$ 3° (H $\times$ V). The camera was aimed such that the target vehicle was approximately in the middle of the image. A selection of the imagery was copied to video disc (sections 3.2.2 and 3.2.3) for use in the laboratory experiments.

3.2.2 Experimental Setup

A flexible setup was developed to present dynamic video imagery to four observers in parallel. The most important properties of the setup are described below.

3.2.2.1 *Dynamic Stimulus Display.*— The heart of the setup was an analogue video disc system (Sony LVR-6000/LVS-6000P) used to present the stimuli to the observers. This system is ideally suited for these kinds of observation experiments for two reasons. First, it allows the use of video sequences as stimuli, which comes as close as possible to the real field operation of thermal imagers because the image dynamics (spatio-temporal noise and image jitter) are retained. Stationary and moving targets can be displayed realistically. Second, it allows the presentation of stimuli (short video sequences) to the observers in random order at fast pace. This is a requirement in the design of the observation experiments. The stimuli were displayed on Sony PVM 122 CE 12-in. monitors white B4 phosphor, and the contrast and brightness controls were set for maximum linear contrast range before the experiment. The observers were not allowed to touch the controls.

The experiments were controlled by a PC that operated the video disc and was further interfaced (RS232) to four response panels used by the observers.

3.2.2.2 *Response Panels.*— Tandy Model 100 notebook computers were used as response panels. A number of keys on the keyboard were designated as target response keys by putting a name sticker on them. Six keys were assigned to the different targets: Leo, AMX-30, PRI, PRAT, AMX-10, and Truck. The camouflaged versions of AMX-30, AMX-10, and Truck were not used as separate response categories. The reason is that interest is in the ability of an observer to identify or recognize a target vehicle either camouflaged or uncamouflaged, not in his ability to distinguish a camouflaged target from an uncamouflaged one.

Three keys were used to further qualify a response as an I (identification), R (recognition) or D (detection only). The latter responses are analyzed in section 5. The LCD display on the notebook computers was used to inform the observers on the status of the experiment and to give feedback during the training sessions.

3.2.2.3 *Observer Setup.*— The observers were placed in a dimly lit room, and the response panel display was illuminated with a small external light source. Care was taken to prevent stray light from falling on the monitor screen. The

observers were allowed to choose their own optimal viewing distance and to scrutinize the display if they wished to do so. The viewing distance was 80 cm on average.

The observers were watched from a control room with a closed circuit TV system. An experiment lasted three days. The sessions lasted 30 to 45 min. The observers worked in two shifts. While four observers were running the trials, the other four had a rest period.

3.2.3 Stimulus Preparation

The video discs can store 24 min of video, or 36000 frames at 50 Hz per side. Because a large number of stimuli had to be stored on a single side, a limited number of frames was available per stimulus. Different sets of stimuli were used for the observer training and for the main experiments. In experiments using only stationary target images from Scenario 1, the observers were trained with stimuli generated from Scenario 2 tapes (see below). In using experiment images from Scenario 1 and 2, about 40 percent of the available images were set aside for observer training.

3.2.3.1 *Stimuli for Experiments.*— For the images of stationary targets, sequences of 2 s (50 frames), taken at each vehicle stop position (section 2), were copied from the Scenario 1 video tapes to the disc. The required stimulus duration was 5 s in the experiments and 4 to 9 s during training. The longer presentation times were obtained by playing the sequences repeatedly back and forth for as long as required. Smooth stimulus presentations of any desired duration could be obtained, and it was not possible to see the difference with continuous video, at least for stationary targets.

For the stimuli consisting of moving targets (Scenario 2), sequences of 5 s duration were copied to a separate video disc. To have the same observation distances as in the Scenario 1 imagery, the stimulus sequences were taken at exactly those times when the targets were passing the stop signs along the track. Section 2 contains details and maps of the approach routes for the two scenarios.

3.2.3.2 *Stimuli for Observer Training.*— The observers were trained in four phases (section 4). For the first two phases, a sequence of close-up images, recorded at 400- to 600-m distances during a characterization session were used to teach the observers the target names and their typical characteristics. Thermal and visible charge-coupled device (CCD) images of all targets, in front and two side views, were copied to the video disc.

In further training stages, the images were similar to those used in the experiments. For experiments where only stationary targets were used, stationary training images were extracted from Scenario 2 (moving vehicles) by copying very short sequences of 15 frames to the video disc and playing them back and forth. The moving targets appeared stationary except at the shortest distances in which case a slight rocking movement was visible. The movement made them a little easier to detect, which was considered an advantage in the training stage. Images of camouflaged vehicles were not used during training.

3.2.4 Stimulus Presentation

Because the present experiments included target identification and recognition only, the image presentations were structured such that the observers could always find the targets easily. Two ways of ordering the target images were used: position and sequential.

3.2.4.1 *Position Presentation.*— In this case, targets popping-up at random positions were simulated by presenting targets at random distances along one of the two approach routes. First a position was chosen at random. To avoid search and detection problems, this position was shown to the observers. All target images available at that position (usually 6 to 8) were presented in a randomly ordered sequence.

All images were presented to the observers five times during different sessions.

3.2.4.2 *Sequential Presentation.*— A straight-line approach of the target toward the MIA was simulated by presenting the targets as an ordered sequence of decreasing distance, from 4 km down. The reasoning behind this presentation

method is that in a practical field situation, targets may often be seen approaching for a certain period of time. As the target approaches, the observer may accumulate information on the target. The accumulation may lead to a better acquisition performance than the targets presented at random positions.

Sequential presentation of the images was repeated three times. Section 4 presents the results of the comparison between the two presentation orders.

3.2.5 Statistical Analyses

The direct observer scores (the numbers of correct responses for all the conditions) are the data analyzed. A correct identification is made if the observer chooses the correct target response key. The response is a correct recognition if target and response belong to the same vehicle class (both are tanks or APCs). Identification and recognition scores were obtained in a single experiment. Most of the results in this and subsequent reports are presented as plots of percent correct responses (identification or recognition) versus target distance. Error bars are shown in some of these plots to give an indication of the accuracy.

The error in the observer scores can be calculated as follows. Let p be the probability that a target in a certain image will be recognized or identified correctly. Suppose that this image has been presented n times.

If the outcomes of the trials do not influence each other, this leads to a binomial distribution with mean value

$$\text{score } (\%) = p * 100\%, \tag{1}$$

and standard deviation

$$\sigma = \sqrt{\frac{p(1-p)}{n}} * 100\%. \tag{2}$$

The standard deviation is highest if p = 0.5 :

$$\sigma_{max}\ (\%) = \frac{50}{\sqrt{n}} \qquad (3)$$

With n = 100 independent observations (responses from 100 different observers that see the image once), the maximum standard deviation in the datapoints will be 5 percent; with n = 20, σ_{max} = 11 percent. In the experiment, each image was presented 5 times to about 20 observers.

Because responses of the same observer cannot be regarded as independent, the maximum standard deviation falls between 5 and 11 percent. In some cases, less than 20 observers (but at least 5) were used. This leads to an increase in the standard deviation by a factor ≤ 2. Therefore, the worst-case assumption is 10 percent $\leq \sigma_{max} \leq$ 20 percent. (Note that the actual values of σ will be much smaller when $p \neq 0.5$.)

Because the numbers of correct responses are distributed binomially, they were analyzed by using a log-linear model. These analyses are similar to Analysis of Variance, which would be used for continuous, normally distributed variables.

3.2.6 Observers

Eight male civilian observers, students from a nearby university, and seven male military observers, (drafted) APC drivers/gunners, participated in the experiment. The military observers had a general military training, experience with a number of military vehicles, and experience with two kinds of thermal imaging systems. If no difference in performance between civilians and military personnel was found, future experiments can be carried out with paid volunteers who are more readily available than military personnel.

All subjects were male and between 18 and 25 years old. They were tested for visual acuity before entering the experiment. For all observers, visual acuity

was better than 1.5 $arcmin^{-1}$. Near vision acuity was tested with the TNO Priegel test. All observers scored better 20 mm^{-1}.

3.3 Experimental Design

The present experiment was designed to measure the effects of different conditions such as target presentation order (section 3.2.4), target motion, target camouflage, approach route, POD, or different groups of observers (section 3.2.6) on acquisition performance. Some general design issues are discussed below; some details are discussed more fully in the following sections.

To draw statistically sound conclusions about performance in many different conditions, images are needed for all target vehicles in all conditions, at all ranges. This is called a complete design. With 9 target vehicles, 15 distances on the left route and 11 on the right route, and 4 different daily recording sessions (PODs), for stationary and moving targets a complete design consists of $(9 \times 15 \times 4) + (9 \times 11 \times 4) = 936$ images, that would have been collected in 72 runs. In practice, it proved not to be possible to record all conditions: only 31 runs of Scenario 1 and 39 runs of Scenario 2 could be carried out. Four additional factors further decreased the number of available images: (1) the target vehicles were not barely detectable or detectable at the first two stop positions of route Right, so nine usable stops remained, (2) some vehicles went off track and got lost, (3) in Scenario 1, some stops on otherwise completed runs were missed, and (4) some images proved to be unusable because of unplanned disruptive events in the field. From Scenario 1, this left a set of 331 images for use in the experiments. The effective design for stationary targets is illustrated in table 7.

Each dot represents a usable target image at a stop position. Missing dots in a sequence represent missed stop signs, and the empty cells in the table indicate sessions that were not recorded at all.

In Scenario 2, the targets drove down the tracks without stopping; therefore, in principle, a very large number of short image sequences of moving targets can be extracted from the video tapes. Because image sequences from

Scenario 2 were selected at the locations of the stop signs of Scenario 1, the complete design for Scenario 2 also has 936 stimuli, of which 480 usable images were available. The effective design for moving targets is similar to the one shown in table 7.

Table 7. Experimental design

	Part Of Day 1		Part Of Day 2		Part Of Day 3		Part Of Day 4	
	Left	Right	Left	Right	Left	Right	Left	Right
Leo 2								
AMX-30								
AMX-30C								
PRI								
PRAT								
AMX-10								
AMX-10C								
Truck								

Because the available image data is so incomplete, the effects of different conditions on acquisition performance had to be analyzed in smaller subdesigns. This means that a smaller number of vehicles (usually about four) for which balanced information is available (targets that drove both approach routes at all possible PODs that can be compared) was used in the analyses. Reducing statistical significance of the results. The problem is hard to avoid when imagery is collected in a large scale, multipurpose field test.

3.4 Observer Training

3.4.1 Training Structure

The observers were trained in four phases. The purpose of training was to bring the observers to about the same level of skill at the start of the experiments, and to avoid contamination of the data by the effects of possible additional learning during the experiments. The images used for training are described in section 3.2.3.2. Observer training took 4 to 5 h of running trials and, with two groups working in shifts, was completed in one day.

3.4.1.1 *Phase 1.*— The observers were shown sequences of four pairs of CCDs and forward looking infrared radar (FLIR) images of the same vehicle: a front view two times and both side views. The name of the vehicle was presented on the response panel display. Each image was shown for 4 s. The observers were encouraged to make notes, and details were pointed out to them. The sequence was repeated four times. Phase 1 lasted about 20 min.

3.4.1.2 *Phase 2.*— The front view thermal images of Phase 1 were shown to the observers in random order at a presentation time of 7 s. The observers responded by pressing a key corresponding to a target, and the response was echoed on the LCD display. Feedback was given by showing the correct target name on the display and by beeping when the response was wrong. This phase was repeated until all observers scored better than 95 percent correct. One session consisted of three presentations of all six vehicles and lasted about 4 min.

3.4.1.3 *Phase 3.*— Images of all six target vehicles at distances of 1001, 1096, 1215, and 1350 m were used. The images were taken from Scenario 1 or Scenario 2, depending on the experiment. Each vehicle appeared 12 times; a complete session consisted of 72 presentations. Presentation time was 7 s, and the response was followed by feedback. Phase 3 was repeated four to five times.

3.4.1.4 *Phase 4.*— Phase 4 was very similar to Phase 3, but closely resembled the real experiment. Target vehicle distances bracketed the complete range up to 4000 m. Images were taken from the left and right approach routes at 4 PODs. A Phase 4 session consisted of 150 stimulus presentations: 2×8 distances $\times$ 6 vehicles from route left and 9 distances $\times$ 6 vehicles from route right. Stimulus presentation was 9 s in the first session. It was decreased to 7 s and, finally, to 5 s in subsequent sessions. In addition to the target response keys, the acquisition level keys, I (identification), R (recognition), and D (detection only), also had to be used. The duration of a session was 30 min, and feedback was given. Phase 4 was carried out four or five times, depending on the scores.

3.4.2 *Confusion Matrices*

A confusion matrix gives a quick impression of the results of an observer on a particular session. The matrix is a table that shows how the responses of the observer are distributed over the different targets. Figures 8a and 8b are two examples of the matrix.

The targets are listed in the left column, and the response categories are listed in the top row. The row to the right of each target contains the percentages of responses in each category given to that target. Consider figure 8a. This is the confusion matrix for one observer after only a few training sessions. The first row of data shows that of all Leopard 2 presentations, 58 percent were correctly scored as Leo 2. However, the Leo 2 was seen as an AMX-30 in 8 percent of the presentations, and it was identified as PRI or PRAT in 17 percent of the cases. The errors are called confusions. The confusions for the other targets can be analyzed likewise. Note that responses on the diagonal (bold) are the correct responses. This observer, in this session, had a total percentage of 69 percent correct.

Figure 8b shows that performance is much better rafter training is completed. Only a few confusions between PRI and PRAT remain.

Confusion matrices were calculated after each session and were discussed with the observers to help them improve their training score.

a

	Responses (%)					
	Tank		APC			Wheel
Test object	Leo 2	AMX-30	PRI	PRAT	AMX-10	Truck
Leo 2	**58**	8	17	17	.	.
AMX-30	17	**75**	.	.	8	.
PRI	8	.	**75**	17	.	.
PRAT	.	17	33	**50**	.	.
AMX-10	33	.	.	8	**58**	.
Truck	.	.	.	.	.	**100**

Figure 8a. Confusion matrix for an observer after a few training sessions. Numbers indicate the percentage of responses assigned to each category with correct responses on the diagonal. Numbers in off-diagonal cells show confusions between targets. The overall correct score is 69 percent.

b

	R	
	Tank	
Test object	Leo 2	AMX-30
Leo 2	**100**	.
AMX-30	.	**100**
PRI	.	.
PRAT	.	.
AMX-10	.	.
Truck	.	.

Figure 8b. Confusion matrix for the same subject as shown in figure 8a after training was completed. Overall correct score is 96 percent.

3.5 Results

3.5.1 Training Results

Figure 9 presents the results of the training for the civilian and military observers and an indication of their performance in the main experiment. The data points present identification scores for individual observers, averaged over target type, approach route, and POD. Three kinds of data are presented: a) the results of Phase 3, sessions 1 through 5 (connected points); b) the average score for a subset of Phase 4 trials, indicated as session 6; and c) the average score on a subset of the main experiment trials, indicated as session 7. Details on these subsets are given below.

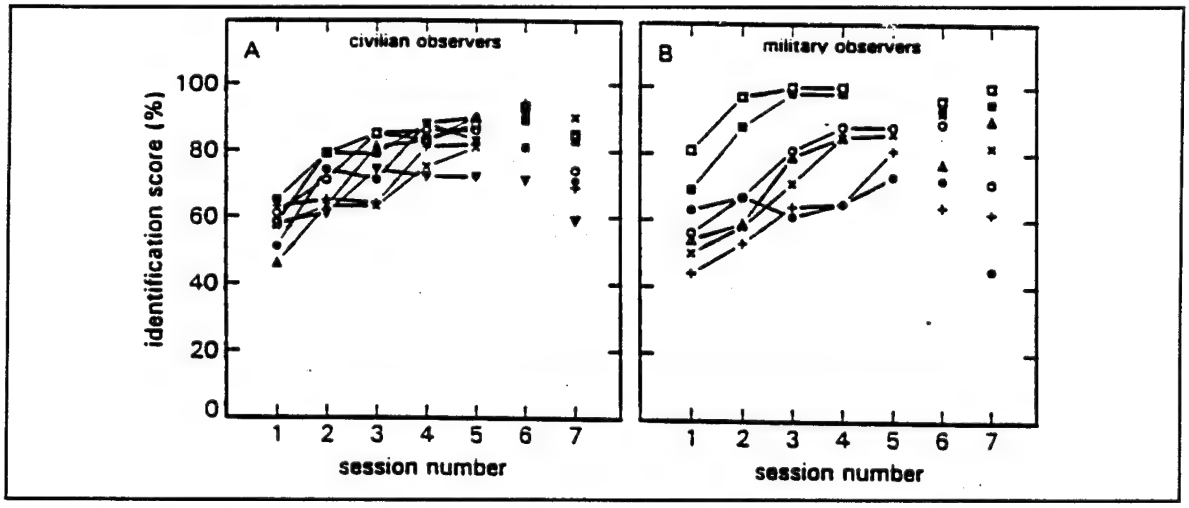


Figure 9. Results of observer training (a) civilian observers; (b) military observers. Three data sets are shown in each panel: score as a function of session number for Phase 3 (connected symbols, sessions 1 through 5), mean score for Phase 4 (number 6), and mean score for the main experiment (number 7).

3.5.1.1 *Phase 3 Results.*— The general trend in the Phase 3 results is that performance increases with session number, which means that the observers are learning. Most observers have reached a steady level after about 5 sessions, which indicates that no further learning occurs.

Figure 9a (civilians) shows that four observers are good (have reached a high level at the final session of Phase 3), three observers are reasonable, and one observer is weak. Figure 9b (military) shows that two observers are extremely good, three observers are reasonable, and two are weak. The civilian observers show rather similar behavior; whereas, there are large differences between the military observers. This difference can possibly be explained by the fact that the civilians, being students, were more homogeneous as a group and were used to sitting down at a task for long periods of time. The military observers (draftees), on the other hand, were not a homogeneous group with respect to educational level, and they exhibited large differences in attitude and motivation.

3.5.1.2 *Phase 4 Results.*— The second data set in figures 9a and 9b (session 6) shows the average scores for the shortest target distances (ranges used in Phase 3) of the Phase 4 trials. For almost all observers, the scores are roughly at the end-level of Phase 3, meaning there is sufficient transfer of training from Phase 3 (only short ranges) to Phase 4 (ranges up to 4 km).

3.5.1.3 *Main Experiment Means.*— The third data set in figures 9a and 9b (session 7) shows the average scores obtained for the shortest target distances (ranges used in Phase 3) in the main experiment. Most observers, except the weak ones, have a score that is only slightly lower than in the training. This means that there was sufficient transfer from training (with feedback) to the main experiment (no feedback). The slight drop in performance may be explained by different images used in the main experiment, and picture recognition (see below) was not possible.

3.5.1.4 *Picture Recognition.*— During the training process, some observers learned to recognize the pictures rather than the targets on the images. This was possible because of the feedback given after each target presentation. Because target recognition is the purpose of these experiments, picture recognition can contaminate the data. The main reason for using different image sets for training and the main experiment was to avoid the effects of picture recognition. Because no feedback is given in the main experiment, picture recognition cannot lead to higher scores.

3.5.1.5 *Observer Selection.*— The large differences in overall performance between observers brings up the question of observer selection. In psychophysical experiments it is generally not allowed to exclude observers that behave differently from the analyses because it gives rise to biased conclusions. In the present case, the situation is different. The final purpose of these experiments is to evaluate TA models. These models should describe the performance of military observers in the field, and such personnel have usually gone through extensive training. It appears that not everybody can learn the task well enough to be an operator in the field; in practical military training, about 30 percent of the trainees fail. For this reason, a performance criterion was set (in advance) and the worst observers were dropped from the final analyses.

The criterion is based on the following considerations: 1) At the shortest ranges (1000 to 1350 m), recognition is a relatively easy task; an observer that cannot distinguish a tank from an APC at those ranges with a reliability of at least 90 percent is in fact not fit for the task. 2) During the BEST TWO field trials, real time observations by military observers yielded identification and recognition scores at short ranges of 75 and 95 percent correct, respectively. After consulting military experts, it was decided that identification and recognition performance in the main experiments, averaged over all conditions and for short target ranges only, should exceed the following limits:

identification:	better than 70 percent correct
recognition:	better than 90 percent correct

These criteria resulted in dropping two civilian and two military subjects (which is about 30 percent).

3.5.2 Military Versus Civilian Observers

Acquisition performance was determined for 15 runs. In figures 10a, b, c, and d, identification or recognition scores are plotted as a function of target range for four of the runs. The standard deviation (not plotted) in each data point is about 10 to 20 percent. Only the results of the position presentation (section 3.2.4) are shown. Solid lines represent performance for the civilians; dashed lines represent performance for the military observers. In all four

examples, the groups show very similar, if not identical, behavior. Second, the performance/distance relations are very different in the four examples. Because the scores for the two groups are so similar for such widely different conditions, it is concluded that no difference exists in overall performance between well-trained military and civilian observers. Statistical analyses of the data shows no significant main effect on observer groups, meaning that, in further analyses, the results for all observers can be taken together to reduce the statistical errors.

3.5.3 General Observations

The following general observations can be made. Target F on the right approach route (figure 10a) is identified correctly up to 3500 m. The combination of the thermal signature of the target and the local background is apparently such that it stands out clearly, almost to the end of the range. Identification of target A, on the other hand, (figure 10b) starts to become difficult at ranges greater than 1500 m and is down to chance level at 2500 m. The general shape of this performance curve is expected: a gradual decrease in score with increasing range, albeit somewhat steep in this case. A similar behavior is found in figure 10c.

The curve in figure 10d (target F at the left approach route) is very different from the curves in figures 10a, b, and c, but it is similar for both groups of observers. The local background on this left approach route at 2000, 2600 to 3000, and 3600 m is apparently such that this target is extremely difficult to identify or even recognize at those ranges, while at a range of about 2400 or 3900 m it appears to be no problem at all.

A similar behavior was found for many different runs, and because it is caused by a strong interaction between target signature and local background, this behavior is termed target-terrain interaction. Section 4 treats this phenomenon in more detail.

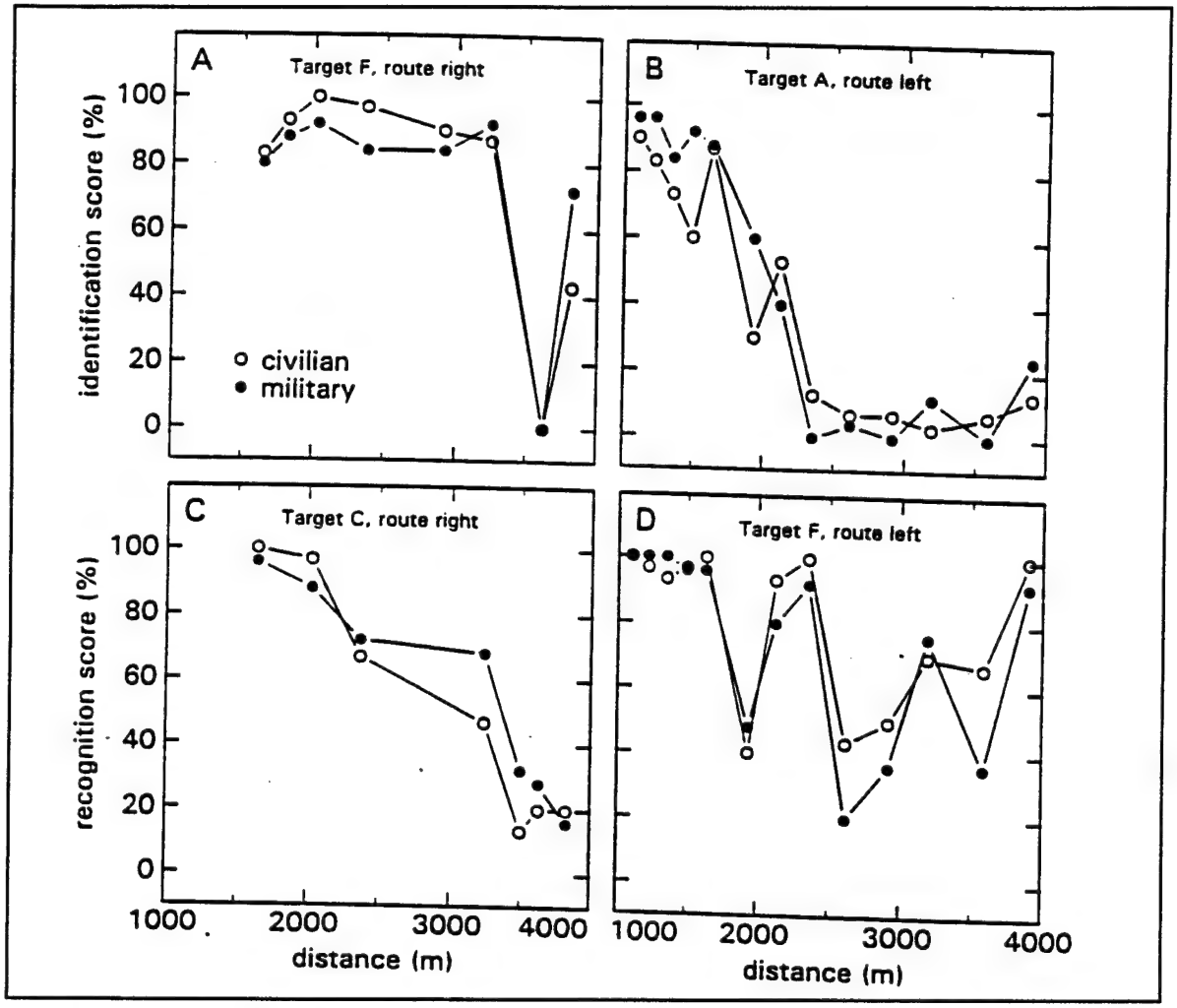


Figure 10. Acquisition performance for civilian (open circles) and military (filled circles) observers: (a) identification score for target F, route right; (b) identification score for target A, route left; (c) recognition score for target C, route right; (d) recognition score for target F, route left.

3.6 Discussion and Conclusions

The experimental setup built for these observer performance experiments proved to be a very practical and flexible tool. Training and experiments were set up quickly and run very efficiently. Because of problems specific to large-field trials, the design was incomplete from a statistical point of view. This problem was alleviated by using many observers and by repeating each experiment several times, reducing the experimental error in individual

datapoints to about 10 percent, which is sufficiently small for our purposes: the evaluation and development of TA models, and the deduction of rules of thumb. Because TA models describe general behavior, only large effects or the absence of effects are of interest.

The method of training the observers worked very well. Analyses of the performance of the military and civilians observers revealed no relevant differences. Note that this finding applies to this study only, where both groups went through exactly the same training procedure. The observer training was sufficient, and the task could be learned relatively easily.

4. Study I: Observer Target Acquisition Performance

4.0 Summary

During BEST TWO, images of single stationary targets (Scenario 1) and moving targets (Scenario 2) were recorded with a thermal imager. The images are used in observer performance experiments to collect data for the evaluation and development of TA models and for operational purposes.

The influence of the head-on motion of targets and the differences in acquisition performance for morning, afternoon, and early night recordings were studied. In a number of situations, performance was heavily influenced by the interaction between the target and the local background properties (target-terrain interaction). Differences in performance for pop-up targets and approaching targets were found. Such differences are not described by current TA models.

4.1 Introduction

The design of the experiments, the setup that was built, the observer training, and the subsequent observer selection process are described in section 3; the main results are presented in section 4, and the effects of several parameters on acquisition performance are determined. A complete set of observer response data is presented in appendix A. The experiments were restricted to target identification and recognition; target detection and search are not studied in the experiments.

The targets are indicated by the letters A through I because recognition performance data on these targets is confidential.

4.2 Methods

The experimental method is explained in detail in section 2. Briefly, a selection of the thermal imagery recorded during the field trial was shown to

observers by using an analogue video disc system (Sony LVR-6000/LVS-6000P). The experiments were controlled by a PC that operated the video disc and collected the responses from a maximum of four observers.

During BEST TWO, recordings were made of stationary (Scenario 1) and moving (Scenario 2) single target vehicles at a range of distances between 4000 and 1000 m. About 10 to 15 short image sequences were taken from each run. A total of 24 Scenario 1 and 14 Scenario 2 runs were used in the laboratory experiments. The images contained one of the nine single target vehicles listed in table 2. Three of the vehicles were camouflaged. The images were presented to the observers for 5 s. The observers tasks were to name the target, and to hit the designated key on a response panel after each presentation.

The analyses of the observer responses are explained in section 3. The results are presented as plots of percent correct responses (identification or recognition) versus target distance. In some of the plots, error bars are shown to give an indication of the accuracy. See section 3 for the calculation of the error bars. In most cases, the standard deviation is 10 to 20 percent.

Two ways of ordering the target images were used: position and sequential. In the position presentation, targets were presented at random distances along one of the two approach routes. In the sequential condition, the targets were presented as an ordered sequence of decreasing distance, from 4 km down. As the target approaches, the observer may accumulate information on the target. This accumulation may possibly lead to a better acquisition performance than if the targets are presented at random positions.

The experiments were preceded by extensive training. The purpose of the training is to bring the observers to about the same level of skill at the start of the experiments and to avoid contamination of the data by the effects of possible learning during the experiments. Training material was selected from runs that were not used in the main experiment. The training method and results are discussed in detail in section 3.

4.3 Experimental Design

The main purpose of the experiments is to collect data for the evaluation and development of TA models. For a thorough evaluation, it is important that acquisition performance is tested under a wide variety of conditions for different times of the day, different atmospheric conditions, different terrain conditions, etc. A second demand, especially important for model development, is to find out which parameters significantly influence acquisition performance. Therefore, it is necessary to vary one parameter, while keeping all other conditions unchanged. To draw statistically sound conclusions about performance in many different conditions, we need images of all target vehicles in all conditions. This is called a complete design. As was shown in section 3, the available material is far from complete, meaning the effects of different parameters on acquisition performance have to be analyzed in smaller subdèsigns, which reduces the statistical significance of the results. Other factors that complicate a comparison between different situations are that background temperature varied during the sessions, and the field recordings were spread over several weeks. The atmospheric conditions were remarkably constant during that period, but terrain conditions changed significantly because of the intensive use of the approach routes.

Research is restricted to a few parameters that seem most relevant to TA modeling: (1) target presentation order (position or sequential), (2) approach route, (3) target motion, (4) camouflage, and (5) POD. The following questions are discussed in this section:

1. Does the presentation order influence acquisition performance? In other words, does accumulation of information lead to higher scores for an approaching target that has been spotted for a while, than for a target that suddenly appears at a certain distance?

2. Is observer performance different for the two approach routes (left, right)? In other words, does local background structure influence performance significantly?

3. What is the effect of head-on target motion on observer performance? In other words, what is the difference in performance between Scenario 1 (no movement) and 2 (continuous movement) images?

4. What is the effect of the camouflage used on three of the vehicles on recognition and identification?

5. What is the effect of POD on observer performance?

The data were obtained in two series of experimental sessions with different groups of observers. The reasons being that not all questions could be answered in a single experiment. About half of the image material had to be used for training purposes. A more complete dataset was collected by using different stimulus material and different groups of observers.

In Experiment 1, acquisition performance was determined for 15 daytime runs (POD 2 and 3) of stationary targets. Both types of presentation order were applied. A group of seven military and eight civilian observers participated in this experiment. As shown in section 3, there is no difference in performance between the two groups. Eleven observers were selected on the basis of a predefined criterion (section 3). The mean scores of the observers are presented.

In Experiment 2, data was collected for 33 runs of stationary and moving targets for all PODs. Part of these runs were also used in Experiment 1. Only position presentation was applied. A new group of seven civilian observers participated in this experiment. After applying the criterion, only four observers were selected and their scores are used in this report. However, two of the observers dropped performed only slightly below the selection criterion. The mean results for six observers were calculated and are very similar to the results for the selected observers, except for a slightly lower overall score. Furthermore, the results for the identical runs in Experiments 1 and 2 are very similar, meaning that, although the scores of only four observers are used in the analyses, the results are reliable.

4.4 Results

4.4.1 Approaching and Pop-Up Targets

In figures 11a through f, identification scores obtained in Experiment 1 are plotted as a function of target distance for six different runs. Open circles represent the scores for position presentation, filled circles represent the scores for sequential condition. The runs are selected to illustrate three different types of acquisition performance versus range found for the position condition (open circles). In figures 11a and b, identification score is invariably high, except for one or two positions, and almost independent of target distance. Apparently, target contrast and camera resolution allow excellent identification of targets F and G, respectively, under these conditions and up to distances of at least 3500 m.

Figures 11c and d show that identification performance is good at short distances, but the score gradually decreases with target distance. For distances beyond 3000 m, the score is at about chance level (17 percent). Such behavior is expected under certain conditions; such as, the resolution is too low to distinguish the various targets from each other at large distances or meteorological conditions limit the identification range (which was obviously not the case during the BEST TWO trials). Notice that the steepness of the curves in figures 11c and d is different.

In figures 11e and f, a very different relation between performance and target range is found. Performance depends largely on the exact position of the target. In these cases, camera resolution is not the limiting factor, because performance is quite good at the largest distance. Similar results were found for three observer groups.

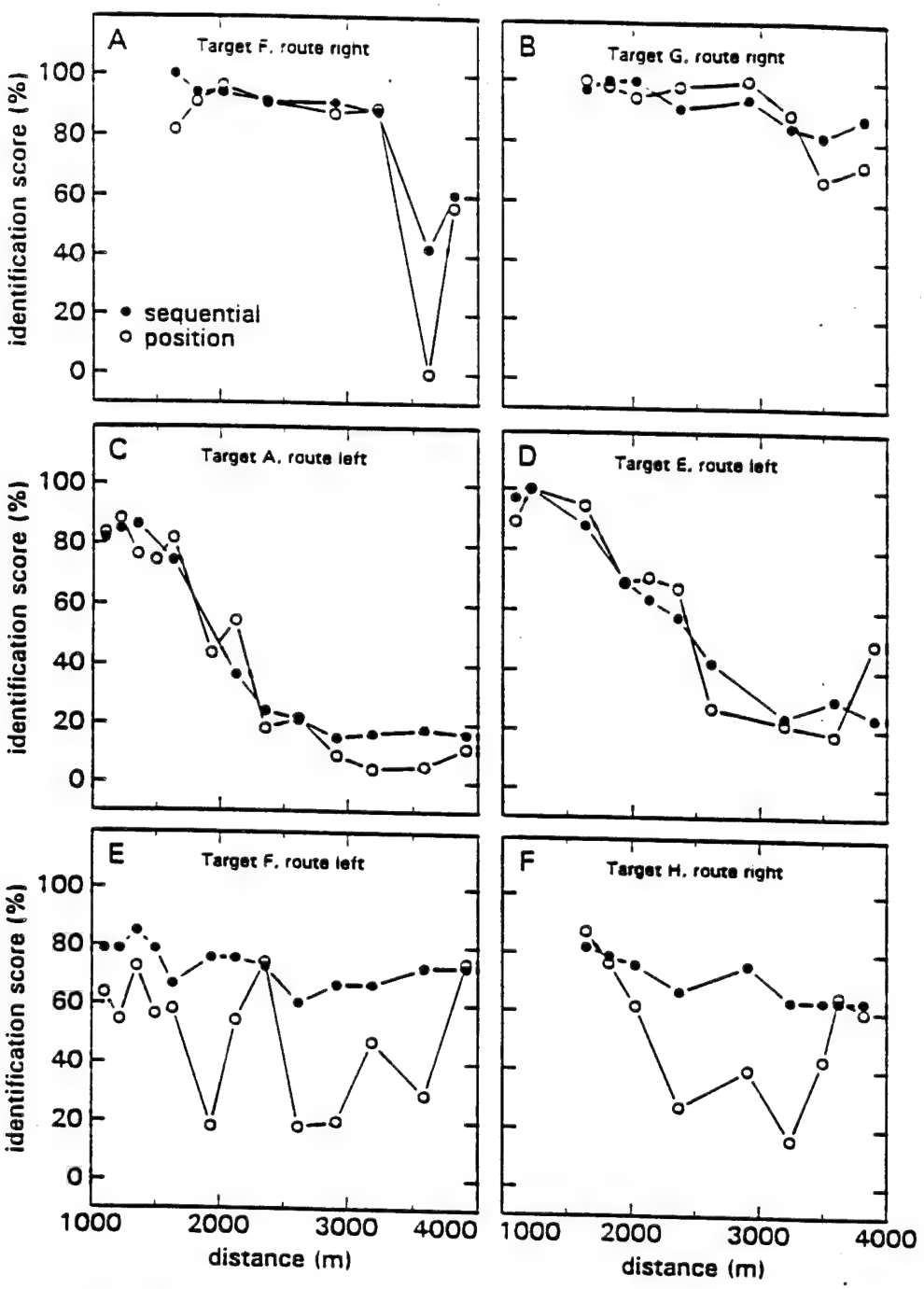


Figure 11. Identification score as a function of target distance for six runs: position presentation (open circles) and sequential presentation (filled circles). For a and b performance is invariantly good; c and d, identification score decreases gradually with target distance; e and f, large target/terrain interactions are found for the position presentation. Sequential presentation leads to more stable results.

In figure 11e the local background on the left approach route at 2000, 2600 to 3000, and 3600 m is such that target F is extremely difficult to identify, while at ranges of 2400 and 3900 m it is quite easy. Inspection of the imagery shows that contrast between target and local background varies greatly during this run: in some cases identification is simple, in others the target is barely visible, and at a few positions it is easily confused with one of the other vehicles. A similar behavior was found for many other runs. Because this behavior is caused by a strong interaction between target signature and local background, this behavior is termed target-terrain interaction.

The filled circles in figure 11 represent the scores obtained with the sequential presentation. In figures 11a through d, the results for sequential and position presentation are similar. Apparently, information obtained from earlier presentations does not improve performance if the information from the actual presentation is equal or better. However, large differences between the curves appear in figures 11e and f. The scores obtained with the sequential presentation are much more stable with respect to target distance. At the positions in which the information is sparse, the observers usually retain their choice from an earlier presentation. As a result, rapid drops in performance, caused by target-terrain interactions, do not take place during a target approach.

Only data obtained with the position presentation will be considered in the remainder of this section. These data contain the most complete information because performance is based on single images without being influenced by history. This makes a comparison between conditions more sensitive to the local properties of target and terrain. On the other hand, if large differences between conditions are found, the differences may be less striking in the case of a target approach.

4.4.2 POD

To determine if the observers perform differently at different times of the day, the effect of the parameter POD on observer performance was analyzed. There are only three conditions for which a comparison between more than two PODs is possible. The recognition scores for these conditions are plotted as a

function of target distance in figures 12a through c. Although the amount of data is too limited to draw sound conclusions, POD does not seem to have a large influence on observer performance. The variability in the data indicates that target-terrain interaction plays a much more dominant role.

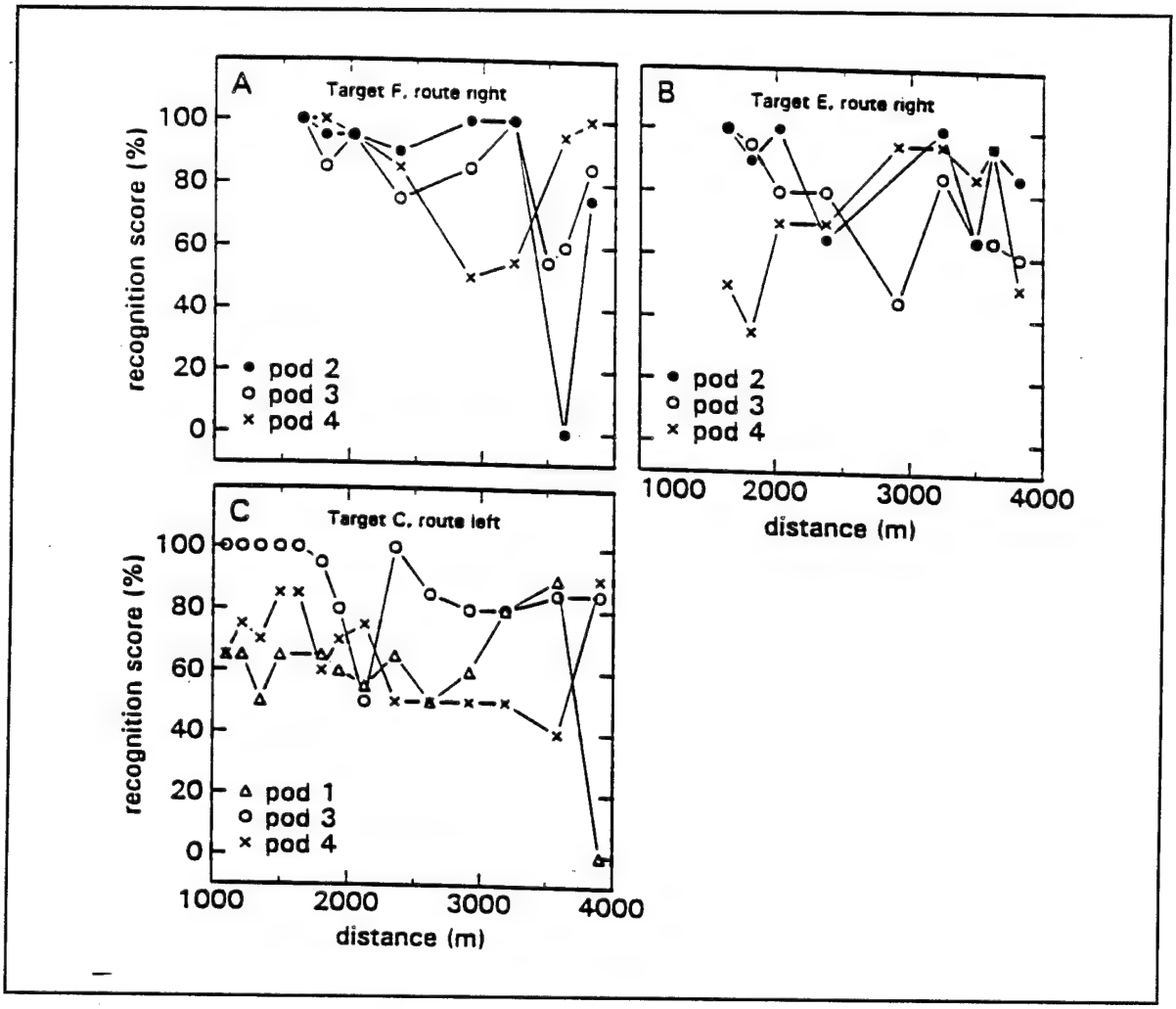


Figure 12. Recognition performance as a function of target distance for different PODs: (a) target F, route right, PODs 2, 3, and 4; (b) target E, route left, PODs 2, 3, and 4; (c) target C, route left, PODs 1, 3, and 4.

A comparison between morning and afternoon (PODs 2 and 3, respectively) is possible for five different conditions (two from Experiment 1 and three from Experiment 2). For two conditions, performance was almost identical; two conditions yielded slightly better scores for POD 2 (figures 12a and 12b), and in one condition the scores for POD 3 were higher. Statistical analyses showed that there is no significant main effect ($P > 0.05$) between the scores for PODs 2 and 3, but there is an interaction between POD and condition. This means that the overall score is similar for the two PODs, although for some runs performance may be different for morning and afternoon.

The differences between daytime (PODs 2 and 3 combined) and nighttime (POD 4) performance were analyzed further. Comparison was possible for six conditions from Experiment 2. Again, no significant main effect was found, but there was an interaction between day/night and condition. The late night (POD 1) was not included in the analyses, because only two late night runs were available in the observer experiments.

The conclusion is that there is no POD for which overall performance is better or worse than other PODs. However, one vehicle may be recognized better during the morning and another during the early night. Such an interaction is not unexpected and may be closely related to circumstantial factors like differences in contrast between target and background.

4.4.3 Approach Route

A comparison between acquisition performance for the left and right approach routes is possible for six conditions: three for which the POD is the same, and three for which it is different. Figures 13a through c illustrate the results of the comparison. Recognition scores are plotted as a function of target distance. Filled circles represent the data for a target on the left approach route; open circles represent the data for a target on the right route.

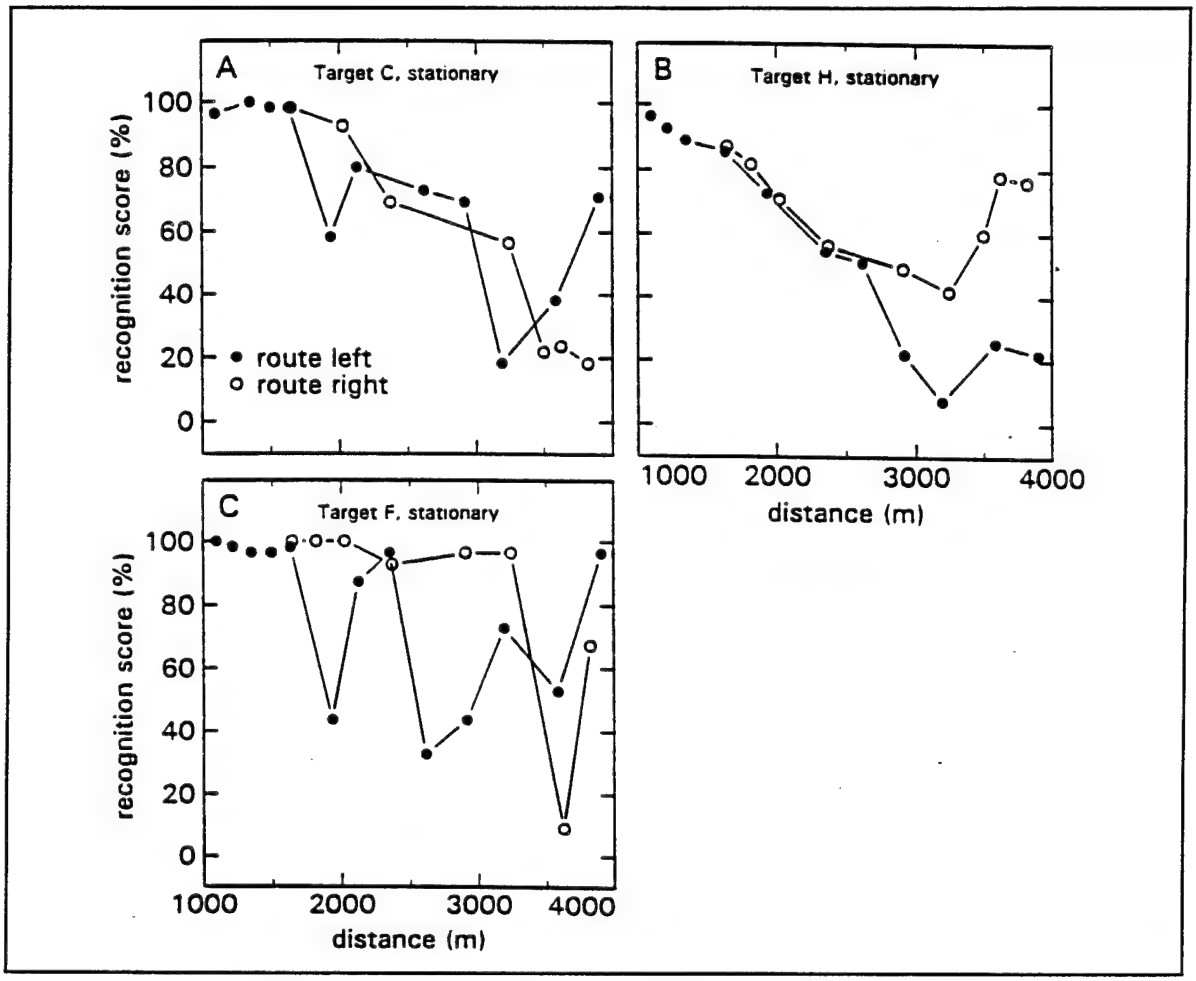


Figure 13. Comparison between acquisition performance for the two approach routes: (a) the overall course of the two curves is similar; (b) performance is identical for the two routes at close range, but significantly better for the right route at large distances; (c) target/terrain interactions cause large performance differences for the two routes.

In figure 13a, the overall course of the two curves is similar, which means there is no important difference in performance for the two routes. Figure 13b shows that recognition performance for the target H is identical for the two routes at close range, but it is significantly better for the right route at large distances. Finally, figure 13c shows that large terrain effects occur on the left route; whereas, target F can be recognized easily on the right route. Thus, depending on the circumstances, performance for two routes, separated by a relatively small distance (60 to 200 m), is sometimes similar and sometimes

very different. One of the reasons for the differences might be that the left route was used more often, and was clearly visible as a white, hot track during the last days of the trials. The results illustrate the importance of the local background structure and its interaction with target signature.

4.4.4 Thermal Camouflage

In figures 14a through f, recognition scores for the camouflaged targets A, D, and G are compared to those for the uncamouflaged versions of these vehicles. Figure 14a is obtained from Experiment 1; figures 14b through f are from Experiment 2. Figures 14a, b, d, and f were recorded on the same day and POD; figure 14c was recorded on the same POD, and figure 14e was recorded on different PODs. Note that the observers were not trained on camouflaged vehicles. Figures 14a and b show that for target A, camouflage does not greatly influence acquisition performance in figure 14a the camouflaged version is recognized slightly better than the uncamouflaged vehicle, and figure 14b illustrates the effects of target-terrain interaction but shows that the overall performance is similar for the camouflaged and uncamouflaged targets. The same findings hold for target G (figures 14c and d). There is a large difference in performance for target D (figures 14e and f). In most situations, target D is recognized very well for ranges up to 3500 m; however, recognition of the camouflaged vehicle breaks down at about 2000 m. The breakdown for target D was found on all material available (four runs). Analyses of the observer responses showed that target D was often confused with target E. Thus, with respect to target recognition, thermal camouflage was not effective for targets A and G, but very effective for target D.

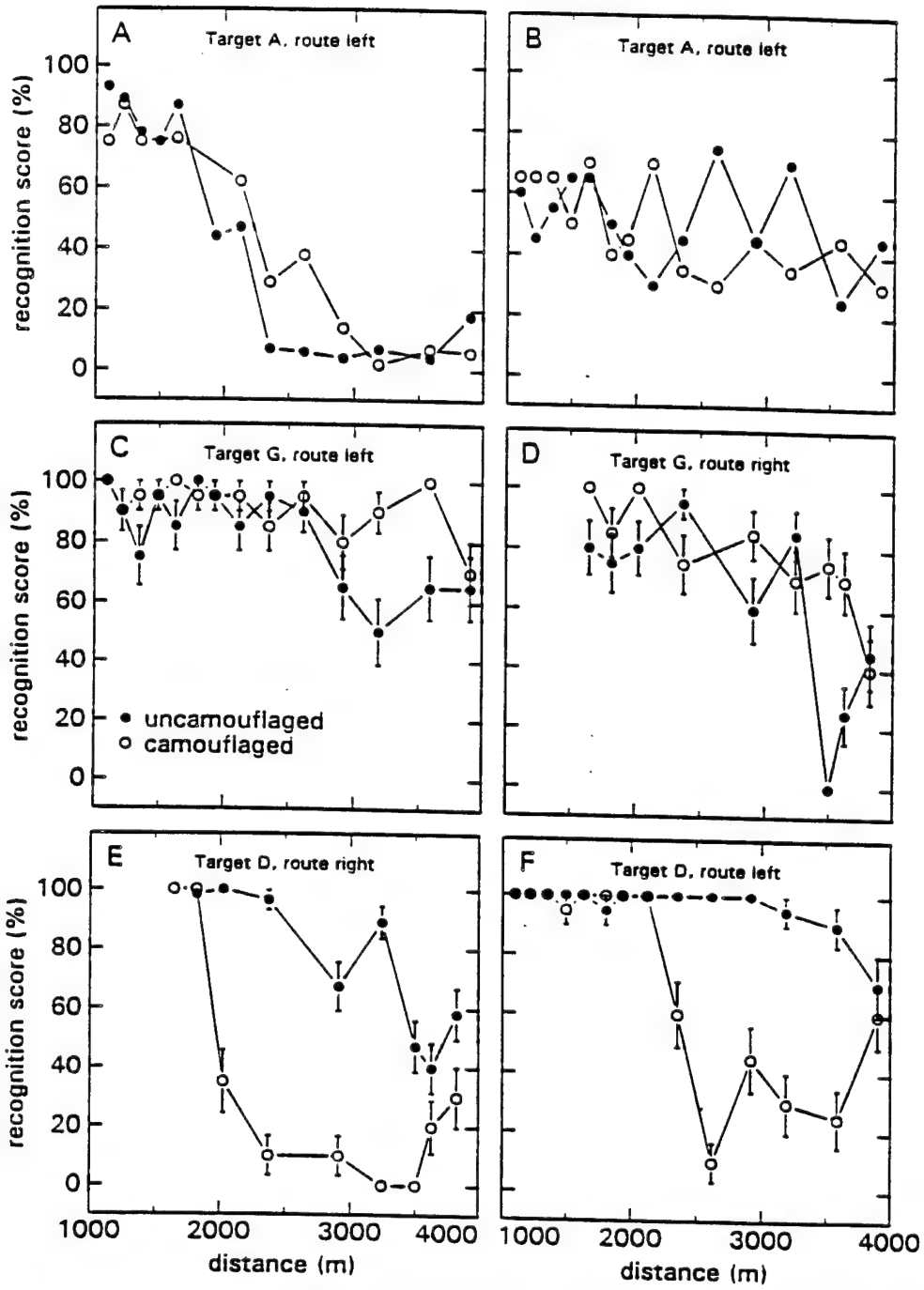


Figure 14. The effects of thermal camouflage: uncamouflaged vehicles (filled circles) and camouflaged vehicles (open circles). No differences in performance are found for target A (a and b) or target G (c and d). The camouflage of target D (e and f) is very effective.

4.4.5 *Target Motion*

A direct comparison of performance for stationary and moving targets was possible for four conditions. The targets were approaching along the left route, and all recordings were made during the mornings (POD 2) of two days.

Figures 15a through d present the recognition scores for these targets, obtained in Experiment 2. Figure 15a shows small differences for target F mainly because of target-terrain interactions. At some positions the moving target is recognized better, at other positions the stationary vehicle is recognized better. The moving target D (figure 15b) is slightly better recognized than the stationary one at long range. A slightly better overall performance is obtained for the moving target G (figure 15c). Figure 15d shows no difference in performance for the stationary and moving target A at large distances; however, the recognition score for the stationary target is much higher at short range.

The differences in identification and recognition performance for stationary and moving targets are small, but statistically significant ($P < 0.05$). The differences are entirely due to the lower score for moving target A at short range (figure 15d): if these data are excluded from the analyses, no statistically significant differences between moving and stationary targets remain. It is uncertain whether the low recognition score for target A is due to target motion per se; other factors may influence performance as well. Inspection of the imagery showed that no dust clouds were near the target. However, the contrast between target and background was much lower for the moving vehicle than for the stationary target, which seriously complicated recognition.

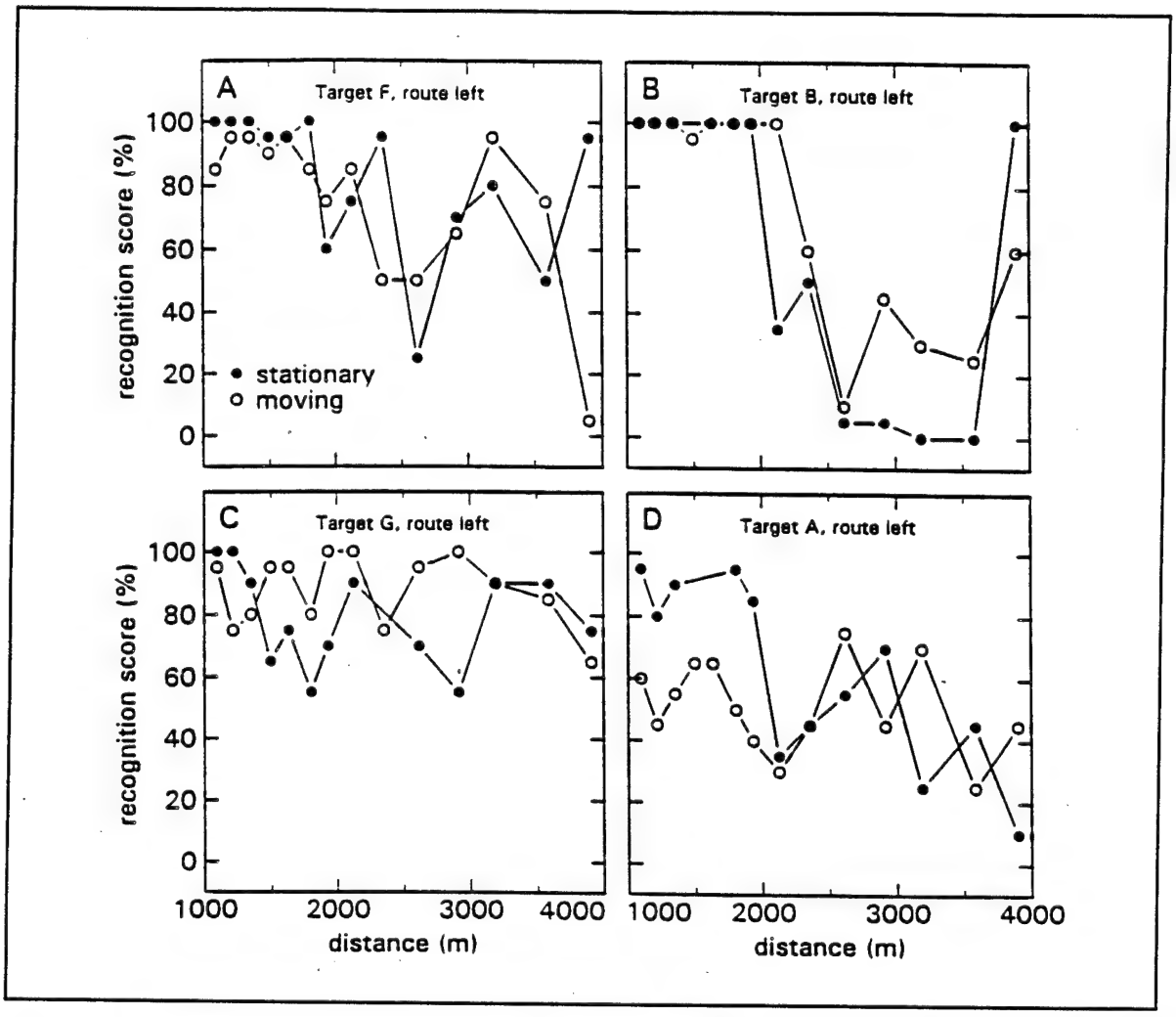


Figure 15. Recognition scores as a function of target distance for stationary (filled circles) and moving (open circles) targets. No overall effect of target motion is found.

4.5 Discussion

Acquisition performance was determined for a large number of runs under various conditions. This section concentrates on the effects of several parameters on acquisition performance.

A major outcome of the experiments is that, in a number of conditions, strong undulations in the relation between target distance and acquisition performance are found. This may be ascribed to the interaction between a target and the

local background properties (target/terrain interaction). A strong target/terrain interaction is found if the contrast between target and background is relatively low, in which case, local changes in background temperature or texture may considerably change the contrast and apparent shape of the target. Terrain interactions complicate the comparison of performance for different conditions. For example, performance for comparable runs along two routes separated by a relatively small distance is sometimes similar and sometimes very different.

Current acquisition models predict a gradual, monotonous relation between target distance and performance, like the curves in figures 11a through d. These predictions are based on parameters like resolution, target size, mean target contrast, and global terrain properties. The models do not distinguish between textured and uniform background. To predict the effects of target/terrain interactions (see figures 11e and 11f), local contrast and local background properties have to be taken into account.

The influence of target/terrain interaction is especially apparent in the results for pop-up targets (position presentation), in which case, the only information available is that of the target at the present position. During a target approach (sequential presentation), information from earlier target positions may be used if the present information is poor. Therefore, the sequential presentation order yields more stable results.

Performance for a target approach can be predicted from the data obtained with the position presentation. The most simple procedure for obtaining an approximation of the score at a certain position, is to take the highest score for the pop-up targets at all positions between the largest and actual distance. In such a model, performance increases monotonically as the target approaches. In a more refined model, the amount of information transfer from each presentation is taken into account to predict the choices of an observer during a target approach. This amount can be derived from the confusion matrix. [6] For example, if the responses to an image are distributed evenly over the response categories, the observer just guesses and the information transfer is low. If, on the other hand, the observer consistently makes the same choice (this choice may be incorrect), the information transfer of an image is high. The image that gives the largest information transfer, contributes most. The

advantage of the second procedure is that dips in performance may occur at certain positions, and it will predict smoother curves than the first procedure. Figure 11 illustrates both phenomena.

There are no indications that overall acquisition performance is different between morning (POD 2), afternoon (POD 3) and early night (POD 4). This means that the amount of clutter, which varied during the day, does not have a large influence on identification and recognition scores. However, performance may be different for individual runs on different PODs.

The thermal camouflage did not reduce the recognition scores for targets A and I, but it was very effective for target D. The camouflage mainly consisted of nets that covered the hot spots like the engine of the vehicles; the shape of the targets was not changed significantly. [7] At the front side of targets A and G, there are no hot spots except the tracks, which explains why camouflage is ineffective for a front-view of these two vehicles.

Head-on target motion did not have a large influence on overall identification and recognition performance. A small (negative) effect was found, but was entirely due to a single run in which the contrast between target and background was very low. A similar finding was reported by Vonhof and Rogge, [8] and by Wester and Van de Mortel, [9] who analyzed observer responses collected during the field trial. Thus, if acquisition models are extended to target motion, head-on motion may be treated as the static situation with regard to identification or recognition. With respect to search and target detection, motion will probably have a large effect.

4.6 Conclusions

1. Observation experiments were carried out with thermal images of stationary and moving single target vehicles. Identification and recognition performance was determined for a large number of runs under various conditions.

2. Head-on motion of targets does not have a large influence on identification and recognition performance.

3. Acquisition performance was similar for morning, afternoon, and early night runs.

4. Performance was largely influenced by the interaction between a target and the local background properties (target/terrain interaction) in a number of runs. These effects are not predicted by current TA models.

5. There is a large difference in acquisition performance for a target that suddenly appears and an approaching target that has been spotted for a while if target/terrain interactions play a role. Performance for approaching targets may be predicted from the results for pop-up targets.

5. Study II: The Reliability of Observer Responses

5.0 Summary

During BEST TWO, thermal images of single target vehicles were presented to observers in the laboratory to determine TA performance. Observers were asked to give two responses after each presentation: (1) they were forced to name the target, even if they were not sure which vehicle was presented; and (2) they were asked to indicate whether they were able to identify (I), recognize (R), or only detect (D) the target. The first (forced-choice) procedure has the advantage that performance is not biased by observer confidence, which appears to yield the maximum score that can be obtained. With the second answer, observer performance is obtained for free—or unforced—identification and recognition reports, which is more similar to the TA task in a practical situation. The scores appear to be partly determined by the observer's confidence. The difference between forced and unforced responses gives a direct indication of the influence of observer behavior on TA performance. The reliability (= probability of correctness) of first, unforced I and R reports during a target approach were also determined. The influence of observer behavior, acquisition level (I or R), target distance, target type, POD, and approach route on these reports was analyzed. The implications for TA modeling are discussed.

5.1 Introduction

Thermal imagery, collected at BEST TWO, was used in an observer laboratory experiment to obtain identification and recognition scores for single target vehicles. The three previous sections describe the experimental method and present recognition and identification performance under a wide variety of conditions. All responses were obtained with a forced-choice procedure (observers are forced to name the target, even if identification or recognition is practically impossible (because the target vehicle is too far away)). The reason for using the forced-choice procedure is that it yields objective scores, which means that performance is not biased by the subjective confidence of an observer. However, the procedure differs fundamentally from the task of a

military observer in a practical situation. If an observer in the field sees an approaching target, he will usually first report a detection, and then, only if he is quite sure, report a recognition or an identification. Because the observer is free to wait until he feels sure, these reports are called unforced. The quality and the number of unforced reports may depend strongly on subjective observer confidence and instructions. If, for example, an observer shows a conservative behavior, targets will be at close distance before he is confident enough to give a report. The reliability of these reports will be relatively high as there are few false alarms. On the other hand, if the observer shows a lenient behavior, his reports will be earlier while the targets are still far away, at the expense of their reliability. Thus, both observer task and observer behavior may influence the outcome of an experiment.

The importance of measurement procedure and observer confidence has long been realized in psychophysical research, and various methods have been developed to separate performance from observer bias. An overview of these methods is given by Bartleson and Grum. [10] However, the influence of these factors on performance has never received much attention. Sanders et al. [11] introduce Signal Detection Theory (one of the psychophysical methods described in Bartleson and Grum [10]) in a recent paper as a tool to determine acquisition performance free from observer bias and re-estimate some of the Johnson criteria. [12] Unfortunately, the experiment was not repeated with the original procedure used by Johnson, which would make possible a direct comparison of the results obtained with the two procedures. Moreover, it is important to keep in mind that observer bias, evidently, plays a role in the practical acquisition task. It is not useful to eliminate this factor by using an advanced psychophysical procedure, if it is not also determined how large the influence of observer behavior actually is. If the effects are large, this may have an important impact on TA modeling.

To simulate the practical situation in the experiment, the observers were not only forced to respond, but they were also asked to qualify each response as I, R, or D. The I and R reports can be regarded as equivalent to unforced identification and recognition reports in the field. Thus, in a single experiment, objective and subjective scores were obtained. The scores can be

directly compared. Furthermore, the effects of differences can be quantitatively determined in observer behavior on TA performance.

The reliability of first I and R reports for approaching target vehicles under a variety of conditions were also determined. A first report is decisive, as in the case that a gunner is instructed to fire as soon as he recognizes or identifies a specific target. A wrong judgement leads to waste of ammunition and may be dangerous. It is important to know the reliability of his initial judgement and to obtain insight on the factors influencing reliability. Such knowledge can also be of use for a commander who may have to make a decision based on several reports from different observers.

5.2 Methods

5.2.1 General

The experimental method is explained in detail in section 3. Thermal images, containing one of nine single target vehicles, were presented to the observers for 5 s. The target vehicles are listed in table 2.

Target distance varied between 4000 and 1000 m. The targets were presented at successive positions, simulating a target approach (a run) along one of the two routes. Table 8 shows an example of a run. In the two leftmost columns, the presentation numbers and corresponding target distances are given. In the first presentation of the run, target distance is 3900 m; in the next presentation, it is 3583 m. A run consisted of 10 to 15 stop positions.

Only daytime runs recorded during the morning or the afternoon of stationary targets were used. The total number of runs was 15. Each run was repeated three times. A maximum of 45 first R and I reports may be obtained for each observer.

Eight male civilian and seven military observers participated in the experiment. Section 2 shows that, 11 of the observers were selected on the basis of a predefined criterion.

5.2.2 *Observer Task*

The observer's task is to name the target and hit the designated key on a response panel after each presentation. In addition, the observer is asked to qualify his response as an I (identification), R (recognition) or D (detection only). If an observer is sure that the presented target was a Leopard 2 tank, he presses Leopard 2, followed by I. If he is sure that it was a tank, Leopard 2 or an AMX-30, he presses Leopard 2 or AMX-30, followed by R. If he thinks it's an AMX-30 (tank) or an AMX-10 (APC), he presses AMX-30 or AMX-10; D, because he is not able to distinguish between two different vehicle classes.

I responses can be regarded as equivalent to unforced identification reports in the field. R responses are equivalent to recognition reports (Leopard 2; R is equivalent to reporting a tank in the field). D responses are not analyzed, because the experiment was designed in such a way that detection of the target was always possible.

5.2.3 *Analyses*

The responses for each run were analyzed in the following way. Table 8 shows a simulated run of a Leopard 2 tank. The two leftmost columns give the presentation number and the target distance, the middle column shows the subject's responses to the successive presentations of the approaching target, the two rightmost columns give the analyses of the responses in terms of a forced and unforced report, respectively.

The Leopard is at 3900 m at the first presentation. The response of the observer, PRI, is of the wrong vehicle type and class. Thus, the forced response is not a correct identification or recognition. The qualification D means that, at this stage, the observer is not confident enough to report an identification or recognition.

Table 8. Simulated run of Leopard 2 tank approaching along the left route

Pres. No.	Distance (m)	Response		Forced Ident./Rec.	Unforced Report
1	3900	PRI	D	wrong/wrong	none
2	3583	Truck	D	wrong/wrong	none
3	3188	Truck	D	wrong/wrong	none
4	2913	PRAT	D	wrong/wrong	none
5	2615	Leopard 2	D	correct/correct	none
6	2353	AMX-30	D	wrong/correct	none
7	2124	AMX-30	R	wrong/correct	correct rec.
8	1933	AMX-30	R	wrong/correct	correct rec.
9	1631	AMX-30	I	wrong/correct	incorr. ident.
10	1494	Leopard 2	I	correct/correct	correct ident.
11	1349	Leopard 2	I	correct/correct	correct ident.
12	1215	Leopard 2	I	correct/correct	correct ident.
13	1096	Leopard 2	I	correct/correct	correct ident.

If all responses, regardless of the acquisition qualification, are taken into account, forced-choice scores are obtained. In table 8, identification is correct for distances below 1494 m and, probably because of a lucky guess, at 2615 m. Recognition scores can be obtained in a similar way. Under forced conditions, the Leopard 2 was correctly recognized as a tank for all distances below 2615 m.

The number of correct identification and recognition responses are divided by the total number of presentations averaged over observers and repetitions to obtain identification and recognition probabilities for each image. Sections 3 and 4 discuss forced-choice performance.

No R or I report was made at distances greater than 2124 m, which means that the observer is not sure enough to pass on information about the vehicle class or type. Thus, in the example of table 8, unforced correct recognitions may only occur for distances $\leq$ 2124 m. From this distance down, all R-reports are correct (the Leopard 2 has been correctly recognized as a tank). Similarly, an unforced correct identification may only occur for distances $\leq$ 1631 m. The first I report is incorrect; at nearer distances, all I reports are correct. Identification and recognition probabilities are obtained by dividing the number

of correct I and R reports by the total number of presentations averaged over observers and repetitions.

The first R report occurs at a distance of 2124 m, and is correct in this example. The first I report (1631 m) is incorrect.

The reliability, or percentage correct, of first reports is defined as the number of correct first I and R reports, divided by the total number of first I and R reports.

An R or I report, given at the largest distance (the first position of a target), cannot be treated as a first report, because recognition or identification might be possible at much longer distances.

5.3 Results: First Reports

5.3.1 Observer Differences

For each observer the overall percentage correct of first recognitions and identifications (averaged over all runs and repetitions is determined). Table 9 presents the results.

At both acquisition levels, the mean overall score is about the same: 75 percent versus 80 percent. The overall variation ($\pm$ 1 s.d.) in reliability is 8 percent for identification and 12 percent for recognition. Individual differences may be caused by differences in observer conservatism (some observers wait a little longer before they report an R or I than others do) or in overall skill of the observers. As shown in section 3, there are large individual differences in the scores at short ranges, which were taken as a measure of observer skill.

Possibly, the best observers also score highest on overall reliability. To test this hypothesis, the correlation between the percentage correct of first identifications (from table 9) and the percentage of correct (forced) identifications at short distances was calculated. The correlation is low ($r = 0.28$) and is not significant ($n = 11$, $P > 0.05$). Similarly, no significant correlation ($r = 0.16$) was found between the recognition scores at short

distances and the score on first recognitions. It is concluded that the variation in reliability is caused by differences in conservatism of the observers (section 5.3.2).

Table 9. Overall reliability (percentage correct) of first recognitions and identifications of civilian and military observers

Observers	Recognition (%)	Identification (%)
civilian		
JB	81	85
EM	82	88
LO	79	93
WH	97	72
PBR	67	77
CK	84	79
mean	82 ± 9	82 ± 8
military		
JS	64	71
RC	87	70
WK	70	79
MW	67	83
PH	55	75
mean	69 ± 12	76 ± 6
overall mean	75 ± 12	80 ± 8

The overall score for civilians is higher than for military observers (13 percent higher for recognition and 6 percent higher for identification). Civilians act more conservative than military observers because both groups score equally well under forced conditions (sections 3 and 4).

5.3.2 The Effect of Target Distance and Route

Figure 16 shows the number of first recognitions and identifications as a function of target distance, expressed as the percentage of the total number of

runs. Figure 17 presents the percentage of first recognitions and identifications correct. The results for the left and right route are plotted together in the figures. Largest distance for the left route is 3900 m; it is 3820 m for the right route. As was argued in section 5.2.3, scores for the distances should be treated separately.

Figure 16a shows that most of the first recognitions (53 percent) are reported at the two largest distances. About 75 percent of these recognitions are correct (figure 17a). This means that recognition of targets at a distance of about 4 km or more is quite possible (more than half of the recognitions can be made at distances larger than about 4 km) for the camera and under prevailing conditions. However, the number of first identifications (figure 16b) made at the largest distances is not higher than at short distances. If the largest distances are omitted, most of the first recognitions are reported between 2200 and 3600 m; the number of first identifications is distributed evenly over the entire range. No R report was given at all a few times. No I report was given in about 10 percent of the runs.

Figure 17a shows that the percentage correct of first recognitions gradually decreases with target distance. However, the mean identification score (figure 17b) slightly increases with distance. Calculation of weighted regression lines yields the following results:

- Between 1000 and 4000 m, the increase of the identification score with distance is about 10 percent. The lines are similar for the left and right routes.

- Over the entire range, recognition scores for the right route are about 10 percent higher than for the left route. The mean recognition score decreases from about 90 percent at short distances to 60 percent at the largest distances.

The slight increase of the identification score with distance is considered not relevant. The significance of the decrease in the mean recognition score with distance is tested below.

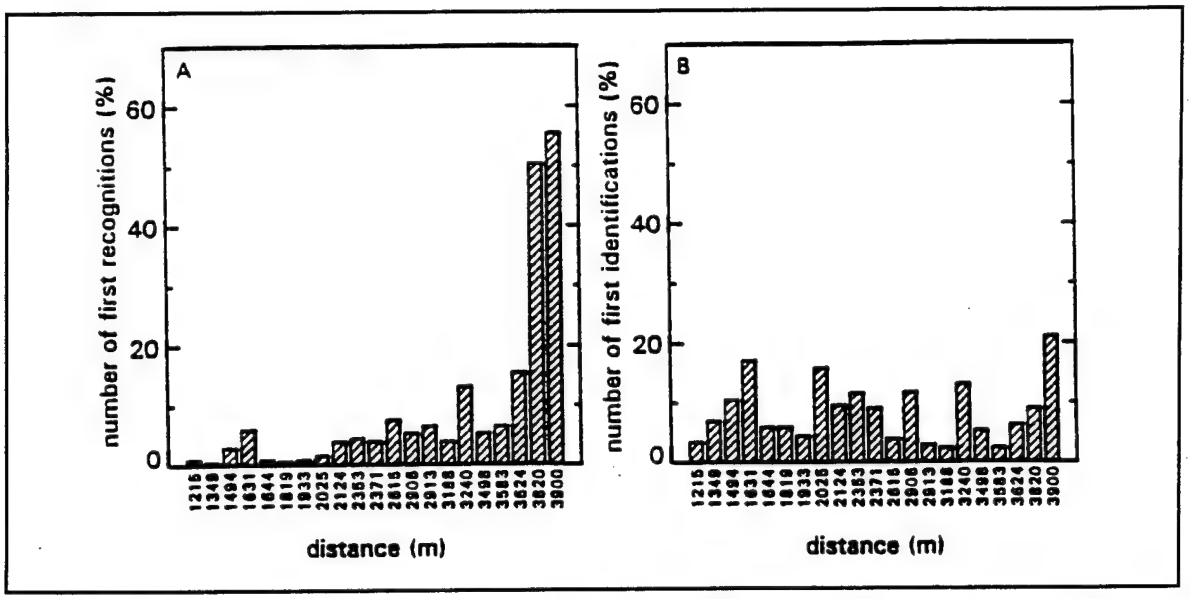


Figure 16. Number of first reports as a function of target distance, expressed as the percentage of the total number of runs: (a) recognition and (b) identification. Most of the first recognitions are reported at the largest distances. The number of first identifications is distributed more evenly over the entire range.

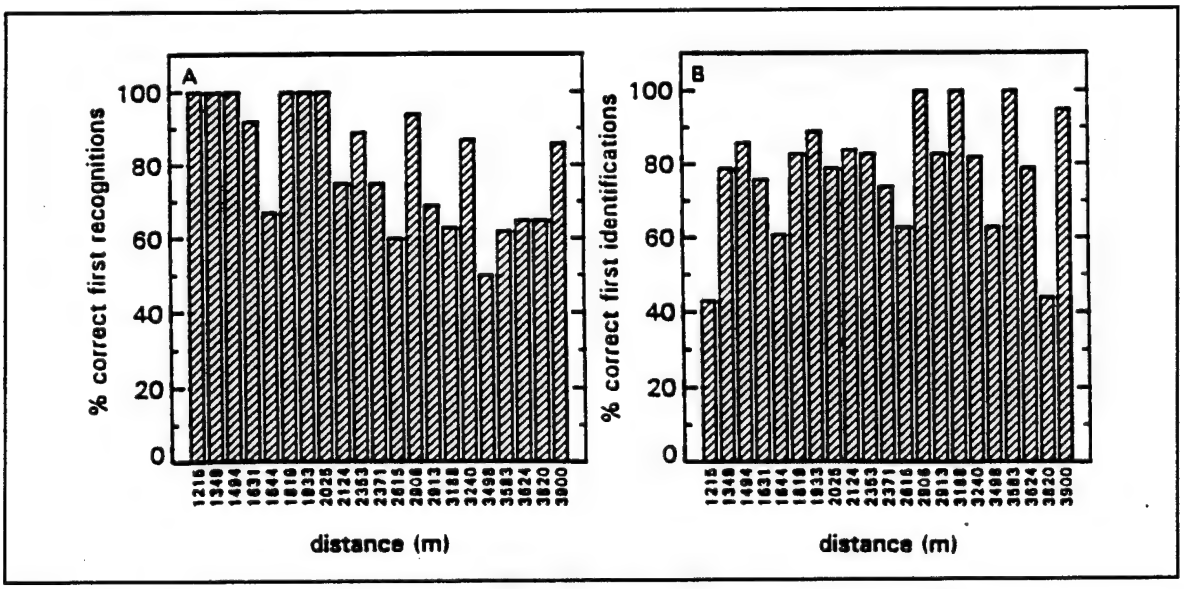


Figure 17. Percentage correct of first reports as a function of target distance: (a) recognition and (b) identification. The percentage correct of first recognition reports decreases with distance. For identification, the percentage is independent of distance.

5.3.2.1 *Interaction Between Observer Behavior and Distance of First Reports.*— The decrease of the recognition score with distance might be caused by a possible interaction between observers and the distance at which the R reports are made: especially for recognition, large intersubject differences were found in mean overall score (table 9).

If two observers are at the same level of skill, the more conservative observer will give his first R report at shorter distances, and yield a higher score. Consequently, higher scores are expected at shorter distances because of differences in conservatism of the observers.

The following procedures were to test whether the decrease in score with distance is real, or caused by differences in observer conservatism:

- The observers were divided into two groups: (1) the recognition scores in table 9 are better than 75 percent (6 observers) and (2) the recognition scores are below 75 percent (5 observers).

- The distances were divided into two ranges: above and below 2500 m to test whether the overall distance effect is present for both groups or is the result of unequal proportions of both groups in the two distance classes.

The test showed that 85 percent of the variation may be ascribed to the observer groups. The remaining 15 percent was caused by target distance and is not statistically significant. The decrease of the percentage correct of first recognitions with distance (figure 17a) is almost entirely due to differential effects of observer conservatism. The test was extended further to determine the influence of approach route (left or right). Differences caused by approach route turned out to be not statistically significant.

It was concluded that target distance and approach route do not significantly influence the reliability of first R and I reports.

5.3.3 *The Effect of POD*

The experiment was carried out for two PODs: morning and afternoon. Mean identification scores are 78 and 81 percent, respectively, and recognition scores are 78 and 69 percent, respectively. An extension of the test, introduced in section 5.3.2, showed that the variation in recognition scores caused by POD is significant ($P < 0.05$) but small, compared to the variation caused by observer differences. It is concluded that there are no relevant differences in reliability caused by POD.

5.3.4 *The Effect of Target Type and Camouflage*

No significant differences in the reliability of first R and I reports were found between target types. However, the results show a slightly lower score (10 to 15 percent) for two camouflaged vehicles than for the uncamouflaged vehicles. The difference is considered not relevant because it is smaller than the intersubject differences.

5.4 Results: Forced Versus Unforced Responses

This section compares the probability of an observer giving a correct identification response under forced-choice to that of giving a correct unforced I report. Figures 18a through d, plot identification scores as a function of target distance for four different runs. Open circles represent unforced responses, filled circles represent forced responses. Roughly, three different cases can be distinguished:

1. Differences between the curves are small if the task is relatively simple (at near distances). In this case, observers are very confident, and most of the responses will be I reports. No differences are expected.

2. Observers become less sure of the vehicle type in a more difficult situation, which means that less unforced identifications are reported. However, identification performance is good if the observer is forced to make a choice, as is shown by the upper curves in figure 18a through c at distances between

2000 and 4000 m. In this case, there are large differences between the scores for the forced and the unforced task.

3. There will be no I reports at all if the task is very difficult (figure 18d). Probability will be at guess level (17 percent for six different response categories) with the force-choice procedure.

For most runs at intermediate distances (case 2), the identification score obtained with the forced-choice task is much higher than it is if the observer is free to respond which means that, although the observer is not confident enough to give an I report, his responses are still quite reliable. Identification ranges which may be defined as the ranges for which identification performance is better than 70 percent may differ by more than a factor of 2 in the examples of figure 18 as a consequence.

5.5 Discussion and Conclusions

Identification and recognition scores were obtained for the TA experiments with two different measurement procedures: (1) the observer was forced to give a response, and (2) the observer was free to wait until he felt sure of the response. The results of these tasks can be compared directly.

A forced-choice task has the advantage that it yields objective scores that are not biased by a subjective confidence criterion of an observer. However, the unforced task is more similar to the actual task of an observer in a practical field situation. Moreover, the chance level, which can be considerable in a forced-choice task with a limited set of targets, has been reduced or eliminated. However, the score depends on a subjective criterion of the observer, which in turn can be affected by the instructions given to him (see below).

The probability that a first unforced R or I report is correct, is about 75 to 80 percent. The percentage is independent of factors that influence acquisition performance, such as target distance, target type, camouflage, POD, or approach route. A similar finding was reported by Vonhof and Rogge, [8] who analyzed observer responses collected during BEST TWO. They reported that the reliability of first I and R reports is about 80 percent, regardless of the

target distance (between 1000 and 4000 m). At the same time, mean acquisition performance varies greatly over that range. Task and camera type used in the experiment were different from those used in this experiment. The results suggest that the reliability of first reports is determined by a fixed internal risk criterion of the observer, which is independent of the outside conditions. An observer will only decide to give a report at a higher acquisition level if he thinks that the risk is acceptable (below his internal criteria). If this hypothesis is correct, the finding will not only be valid for the restricted conditions under which BEST TWO was carried out, but may be considered as a more general rule.

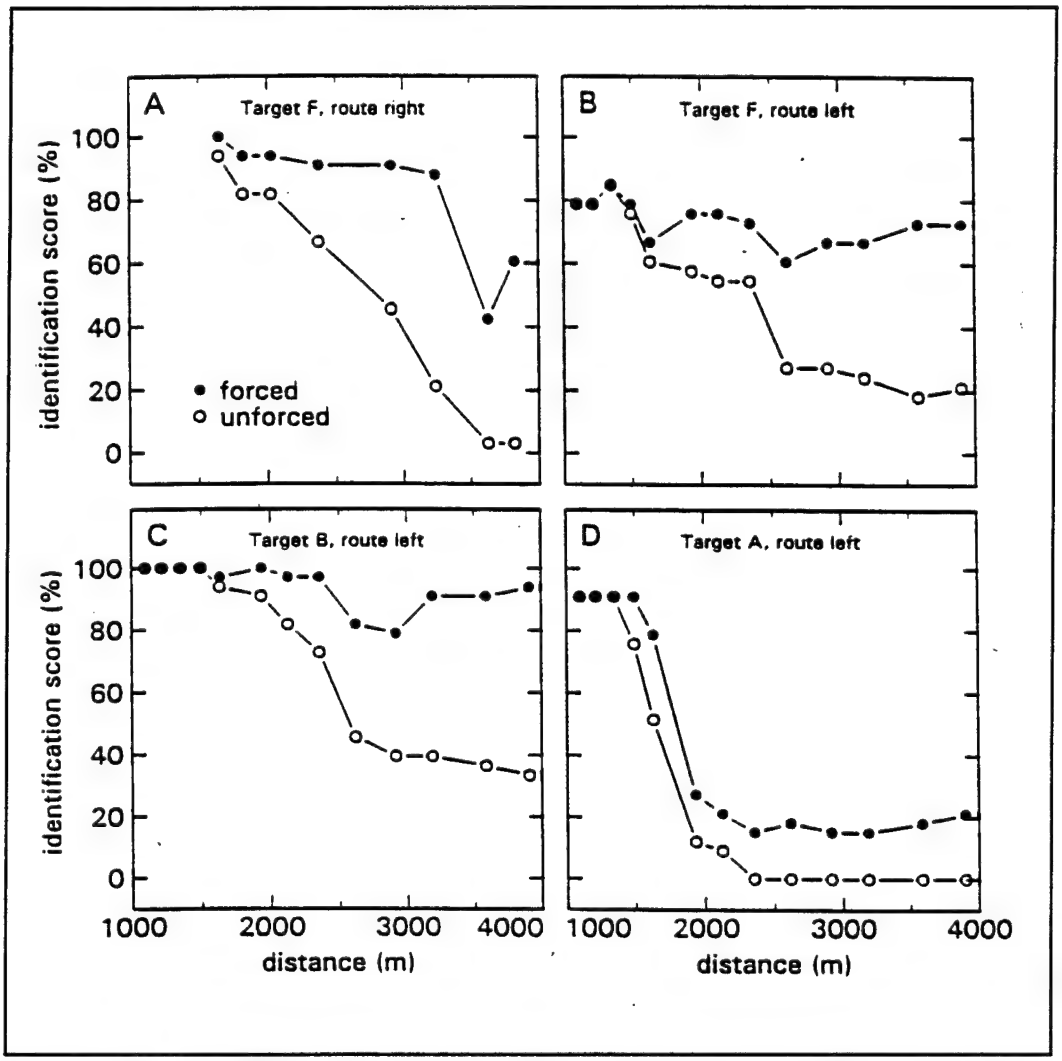


Figure 18. Identification scores as a function of target distance for four different runs: unforced responses (open circles) and forced responses (filled circles).

The present results also show that different observers show a considerable difference in risk criteria, especially for the recognition task: mean correct scores vary between 55 and 97 percent. These criteria are not correlated with observer skill (the ability of an observer to correctly identify or recognize a target if he is forced to). Thus, some observers are more conservative than others, wait longer before they give a report and, consequently, yield higher scores. Surprisingly, civilian observers turned out to be more conservative than military observers.

The two measurement procedures used yield important differences in the probability of a correct response. Identification ranges, obtained with the forced-choice procedure, are usually much longer than those obtained if the observers are free to give a report. This difference may be ascribed to observer conservatism. The largest differences in scores arise in the situations in which the observer is not sure, but chooses correctly if he is forced to. The fact that the differences are considerable means that the observer possesses more information than he actually uses. The risk of making a mistake prevents him from giving a report.

A practical consequence is that the instructions to an observer may have a large impact on the information he will pass on. Therefore, it is advisable that a commander adapts his instructions to the actual situation in the battlefield. If wrong reports are very dangerous, he will instruct the observers to be very conservative, and he will obtain relatively few reports, which will be of high quality. If he wishes to have more information, he will ask the observers to try to identify the targets as soon as possible. This may extend the acquisition ranges considerably, of course at the cost of a higher false-alarm rate. The forced-choice procedure corresponds to the actual limit of unconservative behavior.

The present finding also has important implications for TA modeling. The forced-choice score is the maximum score that can be obtained, given the contrast, resolution, capacities of the human visual system, and chance that an observer guesses correctly. Therefore, the forced condition may be regarded as an important condition for TA modeling. Observer conservatism acts as a

filter that describes the difference between the forced condition and a practical situation.

It is suggested that a TA model consist of two stages: (1) The maximum score that may be obtained under the circumstances is calculated. The score is only limited by the physics and physiology of the systems involved and is independent of observer behavior. Current acquisition models are designed according to this principle. (2) The information is passed through the uncertainty filter, and the actual probability of an unforced report is calculated. The filter characteristics depend largely on observer confidence. Filter characteristics and their variation caused by differences in conservatism can be determined from the experiment. The characteristics also depend on the instructions to the observer. More research is required to determine the impact of this factor.

6. Evaluation of TARGAC Using BEST TWO Observer Performance Data

6.0 Summary

The TARGAC model predicts TA performance for a variety of sensors in the visible and thermal IR. A comparison was made between TARGAC recognition performance predictions and measured observer performance for a large number of trials using thermal imagery collected during BEST TWO. The evaluation shows important differences between measured and predicted recognition performance. On average, observer performance is considerably better than the model predicts: a correction factor of 1.80 should be applied to match the recognition range predictions to the results of the experiments. Further, the model does not give accurate predictions for individual targets on specific backgrounds: the ratio between observed and predicted recognition range varies between 0.9 and 3.6 (95 percent criterion). The routine that describes EO and human visual system performance is responsible for the predictions. This routine is based upon the widely-used NVESD Static Performance Model that uses the Johnson criteria. Hence, the findings of this evaluation may also hold for other models based on these criteria. In addition, it was found that the version of TARGAC tested contained a number of problems and software errors. Corrections are suggested.

6.1 Introduction

TA models predict how well human observers, using an optical or EO viewing device, are able to detect, recognize, or identify a military target. The input variables are the properties of the target and its background, the atmospheric conditions, and the properties of the viewing device used. The output is a relationship between the distance from the target to the sensor and a probability of correct detection, recognition, or identification. TA models are used as TDA's in war games and as a tool to compare performance of competing sensor systems for a specific task. A comprehensive TA model is TARGAC, developed at the ARL-BED. TARGAC is part of the EOSAEL.

The evaluation of TA models is of interest because the reliability and accuracy of their predictions is not always known. The models are usually based on theoretical knowledge of EO device physics, atmospheric optics, and human vision. However, TA is an extremely complex process, and many processes that play an important role in visual performance are not yet understood. Therefore, cognitive factors are often not incorporated in models, and predictions are made for artificial targets in a laboratory environment rather than real targets in the field. The significance of the effects of these omissions are not known. An important side effect is the so-called false precision problem. TA model predictions are often treated as being exactly correct because the accuracy is not known. Therefore, it is necessary to measure observer performance for realistic field conditions and to evaluate model predictions empirically.

Ideally, a TA model evaluation provides a quantitative measure of the accuracy of the model predictions and indicates the applications in which the model may be used and what the restrictions are. The evaluation may also give indications for model improvement. The complexity of the acquisition process makes model evaluation very difficult. It is difficult to obtain accurate and reliable observer performance measures for realistic field conditions. Conditions in the field are hard to control, and there is little or no opportunity for repeated trials under identical conditions, which are needed to obtain statistically meaningful results. Often, only qualitative conclusions can be drawn from the results of a field trial, and the results of the evaluation do not provide insight into the reliability of the model or indicate how the model may be improved. As a result, adequate evaluations of TA models using field data are sparse.

One of the test objectives of the field trial BEST TWO, was to collect observer performance data with sufficient accuracy for a quantitative evaluation of TA models. Many images of stationary and moving target vehicles at many distances were recorded during the test. Thermal (8 to 12 μm) images were used in a laboratory experiment to measure target recognition and identification performance for a large number of observers. A limited observer experiment was carried out in the field for validation of the observer scores measured in the laboratory.

The results of the BEST TWO observer performance experiments are used for the evaluation of TARGAC. A comparison is made between TARGAC recognition performance predictions and measured observer performance for a large number of trials. The result is expressed as a probability distribution of the ratio between measured and predicted acquisition ranges. The mean of this distribution quantitatively shows how well the model predicts overall acquisition performance over a large number of trials. The variance of the distribution is a quantitative measure of the accuracy of the model predictions for individual trials.

The TARGAC evaluation is carried out in five steps:

1. TARGAC predictions are calculated for the BEST TWO situation. To be able to do this, extensive meteorological data for the BEST TWO situation, data on the BEST TWO target set, and the MRTD curve for the thermal imaging system, were collected and fed into the model.

2. Sensitivity analyses are performed because not all the input information is available with a high degree of accuracy. The analyses show the extent to which changes in each input parameter influence the model output. Parameters for which the model is not very sensitive need not be specified with great accuracy, while parameters for which a high sensitivity is found must be provided with high precision.

3. The TARGAC predictions are plotted as probability of a correct recognition response versus target range, with the observer data. Graphical comparison gives a first impression of the quality of the predictions.

4. The ratio between actual and predicted recognition range is calculated for each trial. A set of BEST TWO trials yields a probability distribution of this ratio. Mean and variance of the probability distribution are a measure of the reliability of the model predictions for the set of BEST TWO trials.

5. Further analyses are carried out to find the possible sources of the differences between observer scores and model predictions. Suggestions for model improvement, based on the results, are discussed.

These sections are organized as follows:

- Section 6.2 gives a short description of TARGAC.

- Section 6.3 contains an outline of BEST TWO and the observer performance experiments.

- Section 6.4 presents the sensitivity analyses.

- Section 6.5 gives a simple equation that describes the TARGAC predictions for the entire set of BEST TWO runs.

- Section 6.6 contains the comparison between the TARGAC predictions and the observer performance data.

- Section 6.7 analyzes the variance in the probability distribution to find its possible sources.

- Section 6.8 presents a discussion of the results, and section 6.9 gives conclusions and recommendations.

6.2 TARGAC

6.2.1 General

TARGAC predicts the probability of detection and recognition of military targets as a function of range for a variety of sensors. The model is freely available as part of the EOSAEL in PC and mainframe versions. The program runs in interactive and batch modes. An extensive overview of the model is given in the *TARGAC User's Guide*. [13] The model basically consists of three parts: (1) an inherent target contrast module, (2) an atmospheric effects module, and (3) a system performance routine.

6.2.1.1 Inherent Contrast Calculation.— The first stage of the model calculates the inherent contrast (the contrast at the location of the target) between target and background, given the characteristics of target and background, and the

meteorological conditions. For visual devices, mean or area contrast is defined as the difference between mean target and background luminance divided by mean background luminance. For thermal imaging, target and background temperature are calculated by a Thermal Contrast Model (TCM2), and the inherent contrast is expressed in terms of a temperature difference. It is possible to bypass this module and directly input the inherent contrast.

6.2.1.2 *Atmospheric Effects Calculations.*— TARGAC contains an extensive atmospheric effects module. This module calculates the contrast transmittance through the atmosphere for various wave bands, based on meteorological input data. It yields the apparent contrast of a target as seen by a sensor, as a function of range.

6.2.1.3 *System Performance Calculation.*— The actual probability of acquisition is calculated using the NVESD Static Performance Model [12,14] in which TA performance is described using the well known Johnson criteria. These criteria link TA performance with the ability to resolve dark bars of a certain spatial frequency and contrast against a uniform background. For example, the model predicts a recognition probability of 50 percent if a target is at such a range that a human observer with the viewing system is just able to resolve four line pairs over the effective (minimum) dimension of the target. The higher the resolution of the viewing device, or the larger the target, the longer the range at which four line pairs can be resolved. For a 50 percent detection probability, a resolution of one line pair across the effective dimension of the target is required, for identification this is eight line pairs. Criteria exist for different levels of probability. The relationships between the number of resolvable line pairs and probability for several acquisition levels are called target transfer probability functions (TTPFs). [15] These functions have been established experimentally by averaging over many targets, target orientations and aspects. Ratches [15] also indicates the accuracy of the criteria: the ratio between an optimistic and a conservative criterion is 3:4. The four line pair criterion used in TARGAC for thermal viewing systems for a 50 percent recognition probability is conservative; whereas, a three line pair criterion would be optimistic.

The Johnson criteria are applied in practice using a threshold performance curve of the viewing device that gives the contrast required to resolve a four-bar pattern as a function of spatial frequency. This is called the minimum resolvable contrast (MRC) curve for visible-light devices; IR devices are characterized by a MRTD curve.

6.2.2 TARGAC Version

The TARGAC model is defined as being in the developmental stage of software. [13] This means that new versions are released regularly. For the present evaluation, the PC version of TARGAC that was released in June 1992 was used. A number of software errors were found and most of the bugs were fixed in consultation with Dr. P. Gillespie, ARL-BED, during the sensitivity analyses (section 6.4). This means that results presented in this paper were obtained with an improved version and not with the standard distribution version. Contact ARL-BED before using the program because the standard version still contains a number of errors. Appendix D contains a complete overview of the traced errors, modifications, and recommendations.

6.2.3 TARGAC Input

All calculations were carried out in batch mode for this study. TARGAC requires an input file that contains information about target, background, meteorological conditions, geographic situation, date, time, and the viewing device that is used in this mode. Three levels of probability (between a maximum of 90 percent and a minimum of 10 percent) may be specified in the standard version for which TARGAC predicts a detection and recognition range. The number of levels was extended to five in the improved version. Ranges in this study were always calculated for five probability levels: 90, 70, 50, 30, and 10 percent.

TARGAC has 24 built-in targets and 29 background choices. Examples of targets are T62 and T72 tanks, a Soviet ZIL truck, and a BRDM-2 antitank vehicle. Many of the targets are available in off, idle, and exercised conditions. For example, there is tall grass (growing), dirt road, and coniferous trees (dormant) for background. It is not possible to enter user-specified targets and

backgrounds. The TCM2 module in TARGAC calculates target and background temperature on the basis of target and background characteristics and meteorological input. It is possible to bypass the TCM2 calculations and directly input target and background temperature.

There are 14 visible sights, 4 image intensifiers and 5 thermal sights in the viewing device menu, and the corresponding MRC or MRTD curves are built into the program. The user must specify an MRC or MRTD curve in the form of the coefficients of a sixth order polynomial fit to the MRC or MRTD data when predictions must be made for a viewing device that is not built in.

6.2.4 TARGAC Output

The program calculates detection and recognition ranges for the probability levels specified in the input file. Ranges are specified with a precision of 0.1 km. TARGAC also provides several results of intermediate calculation stages, such as the inherent target contrast when the TCM2 is used.

6.3 BEST TWO Field Test and Laboratory Experiments

The purpose of the BEST TWO filed trials was to quantify the performance of EO devices under battlefield conditions. Section 2 reports a comprehensive overview of the trials.

6.3.1 Observer Performance Data

During the field test, recordings were made of single, stationary and moving target vehicles approaching from 4000 to 1000 m. Targets were always in front view. Image sequences, recorded from a thermal (8 to 12 μm) imager on a U-matic video recorder, were used in laboratory experiments to measure observer performance for target recognition and identification. The experiments are described extensively in the previous sections. For one observer, acquisition performance with the thermal imager was measured directly in the field, and the experiment was repeated with the video tapes in

the laboratory for a number of observers. No significant differences were found between field and laboratory performance. [9]

Observer performance was measured for a total of 38 different target approaches (runs) in two experiments. The runs differ in target type, approach route, date, and time (recordings were made during day and night). Six different targets were used, three of which were camouflaged during some of the runs. Sequences of images containing a single target were presented to observers. The observers' task was similar to the TA task in a practical military situation: after each presentation, they were first asked to indicate whether they were able to identify, recognize, or only detect the target, after which they had to name the target. Two ways of presenting the target images were used: pop-up and approaching.

A randomly chosen target was presented at a random distance in the pop-up presentation. The images from a single run were presented as an ordered sequence, simulating a target approach from 4 km down in the approaching presentation. Search was explicitly avoided. Experiment 1 used 11 observers. Each target image was presented five times. Performance was measured for 15 runs for the pop-up and approaching presentations. Appendix B figures B-1 and B-4 present recognition scores averaged over observers and repetitions for these runs. Each target image was presented five times to four observers in Experiment 2. Performance was measured for 33 runs for pop-up presentation (10 of these runs were also used in Experiment 1). Appendix B figures B-2 and B-3 give recognition scores for these runs.

TARGAC is evaluated against three data sets. Data set A contains observer performance data for all (38) runs for pop-up targets. Data set B contains the data for 15 runs presented as an ordered sequence. Data set C, which is a subset of data set A, contains the data for the same 15 runs now presented as pop-up targets. Data set C is for a direct comparison of the results of the evaluation for the two types of presentation order.

6.3.2 TARGAC Input Data

During BEST TWO, the input data for target acquisition models were collected by several participating nations. Meteorological data were gathered by the delegations of the United States, France, Germany, and The Netherlands. The data are stored in the AAODL database that is maintained by ARL-BED. Dr. P. Gillespie collected the appropriate meteorological and geographical data, date, and time, and composed TARGAC input files (section 6.2.3) for a large number of BEST TWO runs, excluding the viewing device parameters and target and background.

MRTD measurements (horizontal and vertical) for the thermal imager, including the video recorder, were carried out by TNO Physics and Electronics Laboratory (TNO-FEL). [16]

Target and background type have to be selected from the TARGAC menu. Comparable built-in targets had to be chosen (section 6.4.1) because the targets used in the BEST TWO experiments are not part of the TARGAC menu. Direct measurements of target and background temperature in the field, made by the Danish delegation, [17] may also be used. Section 6.5 shows that calculations for the BEST TWO targets can be made using a simplified equation that describes the TARGAC probability versus range predictions for the entire set of BEST TWO runs.

6.4 TARGAC Sensitivity Analyses

All the input data that TARGAC requires to make predictions for the BEST TWO situation is not available, or is not available with high accuracy. Inaccuracies in the input are reflected in the output of the model, affecting the comparison between the model predictions and the observer performance data. The model may be very sensitive to some parameters and insensitive to others. Sensitivity analyses show the extent to which the model outcome is influenced by changes in the input parameters. Basically, this is done by systematically changing one parameter while keeping all the others constant, which allows assessment of the effect of possible errors in the input data on the outcome of the evaluation.

6.4.1 *Inaccuracies in the Input Data*

Inaccuracies in the TARGAC input data for BEST TWO, which may influence the model predictions, follow:

6.4.1.1 *Meteorological Data.*— There are small differences in meteorological data collected by Germany and the U.S. Furthermore, meteorological data are not available for all runs or days. Fortunately, the weather conditions were constant during the trials. Meteorological data from other days may possibly be used in TARGAC. The effects of variations in meteorological data and of using data from other days on the range predictions will have to be assessed.

6.4.1.2 *Target and Background.*— There are two target parameters in TARGAC that affect the range predictions: (1) effective target dimension and (2) thermal contrast between target and background (section 6.2.1.3). The effective dimension of the selected target may differ from that of the target in the field because the BEST TWO targets are not in the TARGAC menu.

The TCM2 in TARGAC and the temperature measurements in the field can provide the inherent thermal contrast of the targets with high accuracy. The model calculates temperatures for the selected built-in target and for a built-in background in TARGAC that do not exactly match the background in the test. TCM2 only calculates mean target and background temperatures; whereas, background temperature varies from location to location. Temperature differences of more than 10 K were found for areas lying only several meters apart in BEST TWO. [17] Field measurements of target and background temperatures may be used, but these were only carried out at one location, and some time before a run. During a run, no measurements were carried out; therefore, measured inherent contrasts are inaccurate. The effects of using a wrong effective target dimension and target and background temperatures on the range predictions have to be calculated.

6.4.1.3 *MRTD.*— The MRTD was estimated in the field through an objective method using an apparatus that was under development. [16] This means that the accuracy of the MRTD may be limited. Bakker and Roos [18] present a large number of objective and subjective MRTD measurements with this apparatus.

On the basis of their results, it is estimated that, with the objective method, the error in cut-off frequency may be up to 20 percent based on the results of Bakker and Roos. [18]

6.4.2 Methods

Section 6.4.1 shows the need to assess the sensitivity of the TARGAC output for variations in time, date, target type (effective dimension and temperature), background type (temperature), and MRTD. Two BEST TWO standard situations were defined as a reference for the sensitivity analyses: (1) Meteorological data were taken from an afternoon trial (Jul 27 at 1500). (2) Data corresponds to an early night trial (Aug 3 at 2300). An exercised T62 tank (TARGAC menu target no. 3) was selected as target, and the background was a grass field (menu background no. 17). The horizontal MRTD, which corresponds to the vertical resolution of the viewing device including the video recorder, was chosen.

Recognition range predictions were always made at 5 probability levels: 90, 70, 50, 30 and 10 percent correct. Apart from the range predictions, target and background temperature calculated by TCM2, also will be considered.

6.4.3 Results of the Sensitivity Analyses

6.4.3.1 *The Effect of Time and Date.*— Figure 19 presents the predicted detection and recognition ranges for the two standard situations. The predicted ranges are almost identical for afternoon and night. Predictions for other days, or other times of the day, yield similar results. TCM2 predicts that inherent thermal contrast is high and only slightly affected by the time of day: contrast is 9.2 K for the afternoon situation and 8.7 K at night. It is concluded that the time of day or date are of minor importance with respect to predicted acquisition range for BEST TWO.

6.4.3.2 *The Effect of Acquisition Level: Transmission Losses.*— The effect of transmission losses through the atmosphere can be assessed by comparing the predicted ranges for recognition and detection in the following way. According to the Johnson criteria, [12] on which the sensor performance model is based,

the number of resolvable line pairs required for target recognition at a certain probability level is four times the number of line pairs required for detection at the same probability level. For example, a 50 percent recognition probability requires 4 line pairs to be resolved across the target, whereas a detection probability of 50 percent requires only 1 line pair (section 6.2.1.3). This means that the ratio between detection and recognition ranges should be 4:1 if atmospheric effects are negligible. Atmospheric effects reduce the ratio.

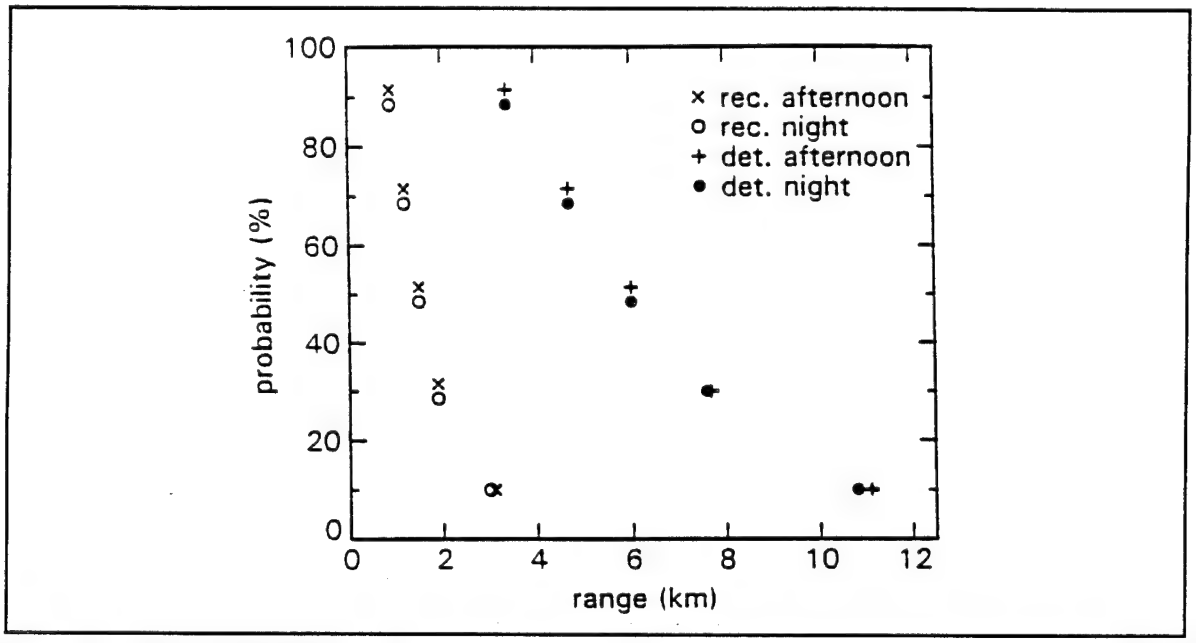


Figure 19. TARGAC detection and recognition range predictions for the two BEST TWO standard situations at 5 levels of probability: 90, 70, 50, 30, and 10 percent. Predictions for afternoon and night are very similar. Where the values coincide, the symbols are shifted slightly in the vertical direction.

Ratios between the detection and recognition ranges for the two standard situations plotted in figure 1 are equal to 4.0 at all probability levels higher than or equal to 30 percent. This means that transmission losses are negligible for ranges below 70 km. The longest target range in the field trial was 4 km. Atmospheric effects do not play a role in the BEST TWO situation, at least if thermal contrast is high.

6.4.3.3 *The Effect of Target Type.*— TARGAC predictions were made for several target types. Targets in exercised and off states were chosen to change thermal

contrast over a wide range. Table 10 presents the results. The first and second column give the target type and its effective, or minimum, dimension. The predicted inherent contrasts for the day and night situation are given in the third and fourth column, respectively. The two rightmost columns present the predicted 50 percent recognition ranges r_{50} for day and night.

Table 10. Effect of target type on TARGAC predictions

Target Type	Target Height (m)	Inherent Contrast (K)		r_{50} (km)	
		Afternoon	Night	Afternoon	Night
T62 (tank), off	2.2	2.6	1.5	1.5	1.2
T62, exercised		9.2	8.7	1.5	1.5
ZIL (truck), off	2.6	3.3	-0.29	1.8	1.8
ZIL, exercised		5.8	2.2	1.8	1.7
T72 tank, off	2.3	2.9	3.6	1.6	1.5
T72, exercised		10.6	11.3	1.6	1.6
BRDM-2 (APC), off	2.1	2.0	1.3	1.4	1.3
BRDM-2, exercised		3.8	3.3	1.4	1.4

6.4.3.4 *Thermal contrast.*— Table 10 shows that thermal contrast varies over a wide range (-0.29 to 10.6 K). The largest differences occur between off and exercised targets. However, thermal contrast only has a small effect on predicted acquisition range. The four predicted acquisition ranges are very similar for each target. Thus, a large influence of target temperature can only be expected for contrasts that are very close to zero (within tenths of degrees), which means that, in most cases, thermal contrast will not be the limiting factor for the predicted acquisition range.

6.4.3.5 *Target effective dimension.*— When atmospheric effects are negligible, predicted range is expected to be proportional to the effective dimension of the target. This is because a number of line pairs must be resolved across the target effective dimension. The minimum dimension is target height for

ground-to-ground TA. The average ratio between the predicted recognition range r_{50} (the two rightmost columns) and target height (the second column) is 0.67 +/- 0.04 for the targets in table 10, which means that the expected proportionality is there. Thus, effective target dimension is a parameter that has a large influence on the model output.

6.4.3.6 *Background Type.*— The test field in Mourmelon mainly consisted of dry grass with bushes. However, because of the frequent use of the approach routes, bare soil came up and hot tracks appeared during the test, which were seen as white lines on the thermal imagery. The targets were seen against a background of wood at the longest ranges. Several background types from the TARGAC menu were chosen to assess the possible influence of background type on the predicted ranges. Table 11 gives the results. Although the temperature is different for different backgrounds (contrast between target and background varies between 12.0 and 5.0 K), there is no effect of background type on recognition range.

6.4.3.7 *MRTD.*— The acquisition threshold will be determined in the high contrast, high spatial frequency region of the MRTD curve, and acquisition range may be expected to be proportional to the cut-off frequency of the MRTD-curve when the apparent contrast is high as for the BEST TWO situation. Acquisition ranges were calculated for the two standard situations using both the horizontal and vertical MRTD-curves for the thermal imager. [16] The results show that the ratio between cut-off frequency and recognition range is constant. Hence, the second parameter that has a large influence on the model output is the cut-off frequency of the MRTD, and an error in this frequency (which may be up to 20 percent, section 6.4.1) directly affects predicted acquisition range.

6.4.4 Verification of the Results for Other BEST TWO Situations

The sensitivity analyses were carried out for the two standard situations only. To check whether the results of the sensitivity analyses also apply to all other BEST TWO situations, TARGAC predictions were made for all the BEST TWO runs for which meteorological data existed. The tank and truck used

Table 11. Effect of background type on TARGAC predictions

Background Type	Temperature (K) (afternoon)	Temperature (K) (night)	r_{50} (afternoon)	r_{50} (night)
deciduous trees	303.5	293.3	1.5	1.5
dirt road, dry	309.0	291.0	1.5	1.5
grass field	304.1	291.9	1.5	1.5
standard sand	304.6	291.1	1.5	1.5
foliage growing sparse	300.6	290.7	1.5	1.5

were the exercised T72 and ZIL, respectively. The BRDM-2 was used for the APC runs, and the grass field was chosen as background. The horizontal MRTD was chosen, which describes the vertical resolution of the system.

Predicted ranges for the same target under different conditions never differ by more than 0.1 km, the precision of the TARGAC output. Thus, the results of the sensitivity analyses apply to all conditions for which there is meteorological data.

6.4.5 Conclusions

The results of sensitivity analyses for the BEST TWO situation can be summarized as follows:

1. Predicted recognition range is almost independent of time or day because of excellent atmospheric conditions.

2. Recognition range is almost independent of target and background temperature because of the excellent atmospheric conditions. Therefore, it is not necessary to have an accurate estimate of target and background temperature. This is a very important result, because large temperature differences were found between different locations in the field, and neither TCM2 nor the field measurements can provide the inherent contrast with high accuracy. It makes no difference whether target and background temperatures calculated by TCM2 or measured during the trials are used for the calculations.

3. There are only two input parameters that significantly influence the TARGAC outcome. Predicted range is directly proportional to the effective dimension of the target and to the cut-off frequency of the MRTD curve of the viewing device. The effective dimension of the BEST TWO targets is known with high accuracy (section 6.5.2). The error in the MRTD cut-off frequency may be up to 20 percent. Thus, a difference of up to 20 percent between actual and predicted ranges may be ascribed to inaccuracies in the values of the input parameters.

4. Only the routine that describes EO and human visual system performance, the NVESD Static Performance Model, can be tested with the BEST TWO observer data because the influences of thermal contrast and atmosphere on recognition range are negligible. This means that the results of the present evaluation may also be relevant for other models based on the Johnson approach.

6.5 TARGAC Predictions for the BEST TWO Runs

Section 6.4 shows that, because of excellent atmospheric conditions during BEST TWO, the predicted probability versus range relationship for recognition depends significantly on only two parameters: effective target dimension and cut-off frequency of the MRTD of the viewing device. Section 6.5 shows that the TARGAC predictions for the entire set of BEST TWO runs can be described with a single equation that contains the two parameters. Such a simplified description of the predictions is convenient for two reasons:

1. The BEST TWO targets are not part of the TARGAC menu, which means that acquisition ranges for these targets have to be deduced from predictions for standard menu targets. Recognition ranges for these targets can be directly calculated with an equation that contains the effective target dimension.

2. TARGAC calculates ranges for only three probability levels (the improved version calculates five ranges, section 6.2.3) in a single run. The entire probability versus range relationship is required for the evaluation (section 6). Calculation of the entire curve would require a number of TARGAC runs for each condition. The derived equation specifies the entire curve.

6.5.1 *Derivation of the Probability versus Range Equation*

Sections 6.4.3.3 and 6.4.3.5 show that recognition range is directly proportional to target effective dimension and MRTD cut-off frequency. The Static Performance Model predicts a recognition probability of 50 percent if four line pairs can be resolved across the effective target dimension (6.2.1.3); therefore:

$$r_{50} = \frac{f_{MRTD} \cdot D_{TARGET}}{4} \qquad (4)$$

where

r_{50} = the recognition range in km at the 50 percent probability level, f_{MRTD} the MRTD cut-off frequency in lp/mrad
D_{TARGET} = the effective dimension of the target in m.

Equation (4) is confirmed by the results of the calculations in section 4 (table 10).

The probability versus range relationship can be described very well with an s-shaped curve known as the Weibull function. The relationship is given by

$$P = \left[1 - 2^{-\left(\frac{r_{50}}{r}\right)^s} \right] \cdot 100\% \qquad (5)$$

where

P = the predicted probability of a correct recognition
r = the target range.

The parameter s determines the steepness of the curve and is set to s = 2.32 for an optimal fit to the TARGAC predictions. Figure 20 confirms that this function nicely coincides with the predictions for one of the standard situations (see 6.4.3).

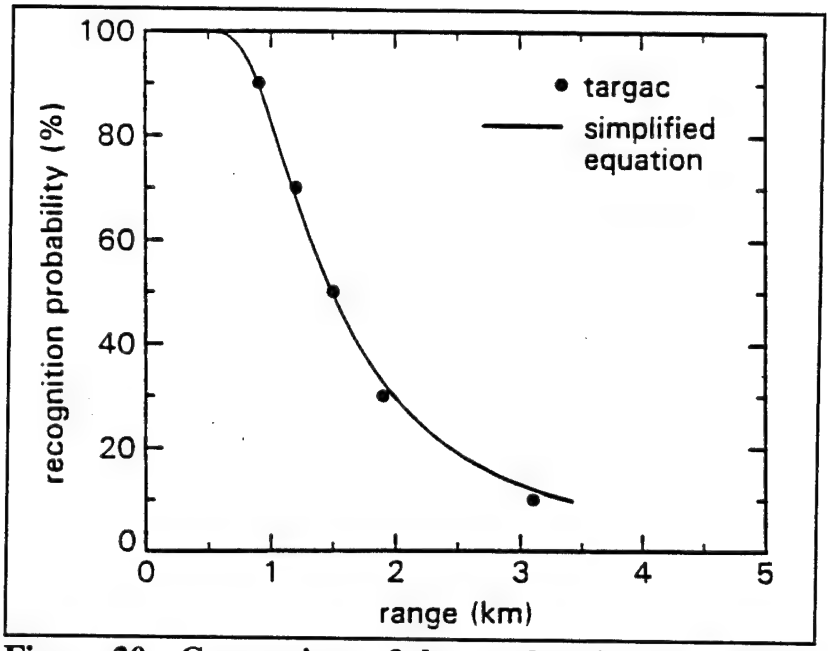


Figure 20. Comparison of the results of the simplified equation (solid line) with the TARGAC predictions (filled circles) for the BEST TWO standard situation.

Equations (4) and (5) can be combined to yield the following equation which describes the entire set of TARGAC recognition range predictions for the BEST TWO situation:

$$P_{TARGAC,BEST\ TWO} = \left[1 - 2^{-\left(\frac{f_{MRTD} \cdot D_{TARGET}}{4\,r}\right)^{s}} \right] \cdot 100\% \qquad (6)$$

or, inversely

$$r_{TARGAC,BEST\ TWO} = \frac{f_{MRTD} \cdot D_{TARGET}}{4} \cdot \left[-\,{}^{2}\log\left(1 - \frac{P}{100}\right) \right]^{-\frac{1}{s}} \qquad (7)$$

6.5.2 Range Predictions for the BEST TWO Targets

Equation (7) can be used to calculate the TARGAC range predictions for the specific targets used in BEST TWO, if effective dimension and the MRTD cut-off frequency of the viewing device are known. Table 12 presents the results.

The BEST TWO targets are listed in the leftmost column. The next columns give the estimate of their effective dimensions and the corresponding 50 percent recognition ranges (r_{50}) calculated with equation (7). The bottom row gives a 50 percent recognition range for a mean BEST TWO target, which will be used in one of the analyses in the next section. Note that the probability versus range relationship is the same for each target and each run, except for a single factor that is determined by the effective target dimension (equation (6)).

Table 12. Target effective dimensions and predicted recognition ranges (at the 50-percent probability level) for the targets used in BEST TWO

Target Type	Target Height (m)	r_{50} (km)
Leopard 2	2.50	1.8
AMX-30	2.30	1.6
PRI	2.60	1.8
PRAT	2.60	1.8
AMX-10	1.90	1.3
Truck	2.80	1.9
mean target	2.45	1.7

6.5.3 Conclusions

TARGAC recognition range predictions for the entire set of BEST TWO runs can be described with a single equation that contains only two input parameters: effective target dimension and MRTD cut-off frequency. Such an equation is convenient for the evaluation of the model because complete probability versus range curves for the BEST TWO targets can be calculated at once.

It is also shown that, after modifications to the software, the TARGAC calculations for the BEST TWO situation are in agreement with the Static Performance Model predictions, as expected.

6.6 Evaluation of Range Predictions

The evaluation of TARGAC takes place at various levels of complexity. The TARGAC predictions for the BEST TWO targets (section 6.5.2) are plotted together with the observer data presented in appendix B in section 6.6.1. This gives a qualitative impression of the accuracy of the model predictions.

The TARGAC predictions are compared to the overall mean score of all runs in section 6.6.2. The Johnson criteria were originally based on mean acquisition performance over a large number of conditions; therefore, it is useful to check their validity in this respect.

Finally, a complete quantitative comparison is made between the set of individual datapoints and the TARGAC predictions. Large differences in performance were found because of factors such as target type, target distance, approach route, and POD in the BEST TWO observer data. The sensitivity analyses showed that the TARGAC model predictions depend on only a few of these factors. Thus, the model cannot account for part of the variation in the observer data. This part, called the unexplained variance, is determined in section 6.6.3. It is of obvious importance to know how large the unexplained variance is, because it directly provides a measure of the reliability of the acquisition ranges predicted by the model. If the amount of unexplained variance is small, the model is able to make reliable predictions for individual cases or trials (a T62 tank on a grass field at 3000 m at 1400). The model is not applicable to individual cases and can only be used to predict overall mean performance if the amount of unexplained variance is large.

6.6.1 Qualitative Comparison for Individual Runs

Appendix B figures B-1 through B-4 present the complete set of recognition performance data from the BEST TWO observer experiments, described in section 6.3.1, with the corresponding TARGAC predictions for these runs. The set consists of 63 plots of recognition performance versus target range. Each plot typically consists of 10 to 15 datapoints. Filled circles represent the averaged observer scores, and solid lines represent the TARGAC predictions.

The standard error of the mean of the observer scores s_p was smaller than 10 percent (see appendix C).

Figure 21 gives three typical examples from this set. The prediction is reasonable: measured and predicted recognition performance gradually decrease with target range in figure 21a. However, TARGAC underestimates observer performance. Predicted recognition performance is far too low in figure 21b. The data show that target recognition probability is better than 80 percent for ranges up to 4000 m; whereas, the predicted probability is below 80 percent at a range of 1000 m. TARGAC predicts a recognition probability of less than 10 percent at a distance of 4000 m. Figure 21c shows that recognition performance does not simply decrease with target range but changes rapidly with the exact position of the target. The recognition probability is high at distances below 1600 m and at 2200, 2300, and 3900 m. At intermediate distances (1900 and 2600 m) recognition probability is very low. This behavior was termed target-terrain interaction (section 4), and it cannot be described satisfactorily with a monotonously decreasing function. On average, predicted performance is far too low.

The following conclusions can be drawn:

- The predicted curves fall below most of the datapoints, meaning the TARGAC predictions are conservative on average. At a certain range, the predicted probability is too low, or equivalently, the predicted ranges for a certain probability level are too small. The deviation can be considerable for individual points. The data often show a high recognition probability at distances near 4000 m; whereas, TARGAC predicts a probability less than 10 percent for these ranges.

- The monotonously decreasing performance curve that TARGAC predicts is found only in a number of cases.

- Strong undulations in the relation between target distance and acquisition performance are found in some cases, because not only target range but also the local conditions are an important determinant of observer performance.

Strictly, a probability versus range relationship does not exist in these cases. Such behavior is not predicted by the model.

- TARGAC is not suitable for predicting performance for individual targets.

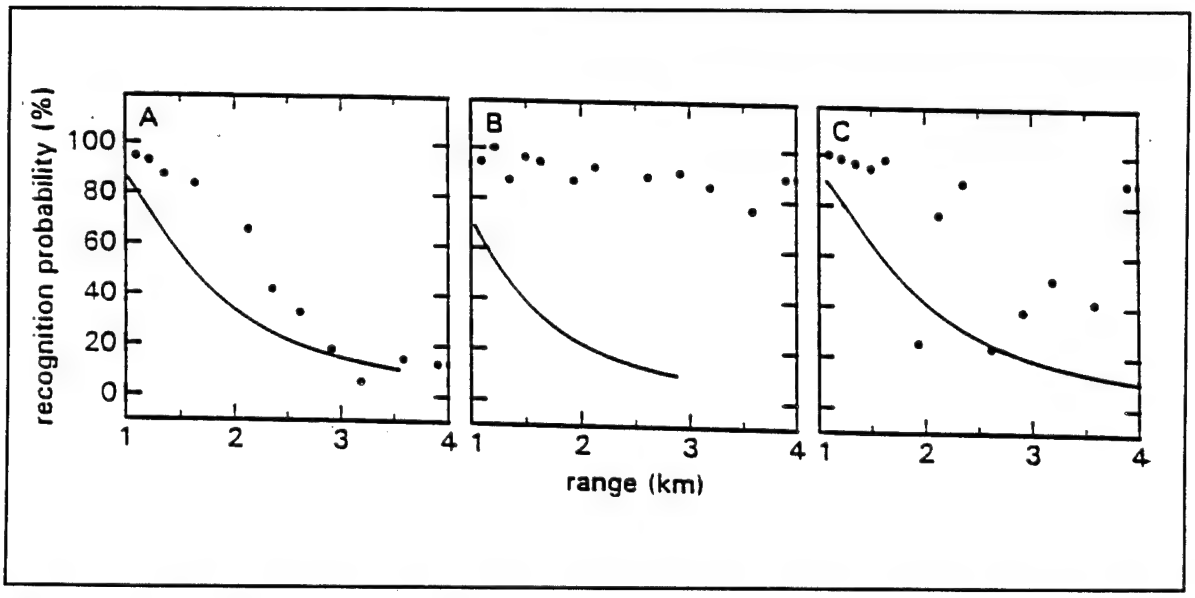


Figure 21. Comparison of observer performance versus target range and the corresponding TARGAC predictions for three typical examples of observer recognition scores (filled circles) and TARGAC predictions (solid lines): (1) measured and predicted performance gradually decrease with target range, TARGAC underestimates observer performance; (2) predicted recognition performance is far too low at all ranges, and (3) TARGAC is unable to predict the large undulations in the observer scores.

6.6.2 Comparison With Overall Mean Observer Performance

Because TARGAC does not appear to be suitable to predict performance for individual targets, the predictions of TARGAC are compared with the overall mean performance of observers over a large number of targets and runs. This seems to be a reasonable approach because, originally, the Johnson criteria were also based on mean acquisition performance over a large number of conditions.

Data sets A, B, and C, defined in section 6.3.1, are used, and figure 22 shows the comparison with the TARGAC prediction for the mean BEST TWO target (see table 12).

Using the prediction for a mean target makes sense because all targets in the set were presented to the observers approximately the same number of times, and predicted ranges for the six targets do not differ considerably. Filled circles represent mean observer performance averaged per distance. The solid line indicates the TARGAC predictions for the BEST TWO mean target. The dashed line represents the best fit of equation (2) (section 6.5.1) to the observer data and is shown to correspond to TARGAC range predictions increased by a factor of 1.80 (section 6.6.3).

Data set A is the largest set. Target images were presented in the pop-up presentation order. Data sets B and C (a subset of A) correspond to the same set of images, but the images were presented as an ordered sequence for data set B, simulating a target approach. Figure 22 shows that

- TARGAC underestimates mean observer performance for all data sets.

- The mean recognition probability for the observers is never below about 50 percent correct, even at the longest target range (4 km). This means that the shape of the measured probability versus range relationship is not known for lower probabilities.

- The overall mean probability decreases with target range, which was not the case for the scores for individual runs (figure 21). This is because averaging over many runs diminishes the effects of target/terrain interactions. The results for set C show more residual terrain interactions because it is a small subset of A.

- Comparison of data sets B and C shows that for approaching targets (B) the data conform much better to a monotonously decreasing function than for pop-up targets (C). This is because the accumulation of information during a target approach helps to reduce the effects of the target/terrain interactions (section 4).

• For the approaching presentation order (data set B), the slope of the dashed curve fits nicely to the data. For pop-up targets, the predicted curve is too steep. A shallower function would fit better. This difference in slope is most pronounced for data set A, which might suggest that the model should be made to accommodate the different target behaviors.

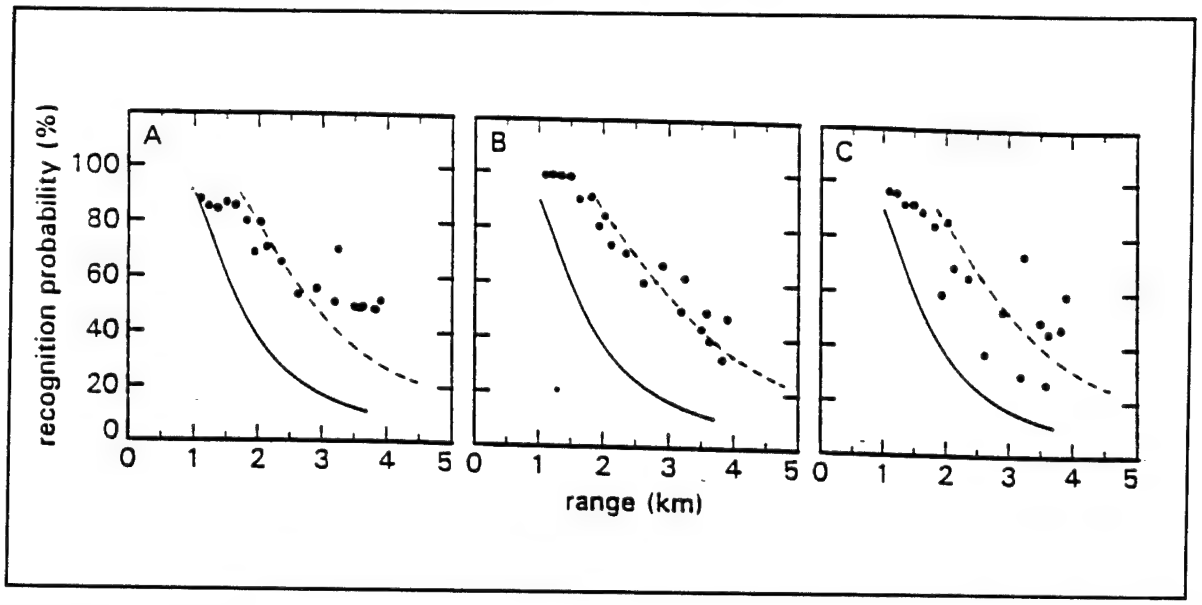


Figure 22. Comparison of overall mean observer performance for data sets A, B, and C, and TARGAC predictions: mean observer recognition scores (filled circles), TARGAC predictions (solid lines), and best fit of equation (2) (section 6.5.1) to the data (dashed line). TARGAC predictions are far too conservative.

6.6.3 Comparison With Observer Performance for Individual Trials

Previous sections, show that overall mean recognition performance can be described reasonably well as a monotonously decreasing function of target range. The fit of the model to the mean observer data could be much improved by applying a single correction factor and, possibly, a small change in the steepness s (equation (3) or (4)). Apart from the mean performance, it is worthwhile to know how well TARGAC predicts the performance for individual trials. The previous sections show much variation in the observer data, but the model predicts only a single curve with an unknown confidence interval.

To determine the unexplained variances in the observer data, the variances will be regarded as a large set of single probability versus range points, and a point-by-point comparison will be made between actual target range and the range predicted by the model (range comparison). These analyses will yield a distribution that gives the unexplained variances in the data, caused by all parameters that were varied in the experiments including the effects of local conditions. The variance provides an indication of the quality of the model predictions for individual trials.

6.6.3.1 *Procedure.*— Each datapoint represents a probability of correct recognition P for a target at range r. At this probability level, the model predicts a target range r'. A ratio r/r' between actual and predicted range is calculated for each datapoint in the set. The ratio $r/r' = 1$ if the model makes a correct prediction. The ratio will be larger than 1 if the predicted range is too short; the ratio is smaller than 1 if the predicted range is too large. Figure 23 gives an example of the procedure for a few datapoints. It is convenient to transform the values to a log scale because correct predictions are centered at 0, and over and underestimates of the range by the same factor are equally shifted in opposite directions along the axis. For example, the predicted range is twice or half the actual range.

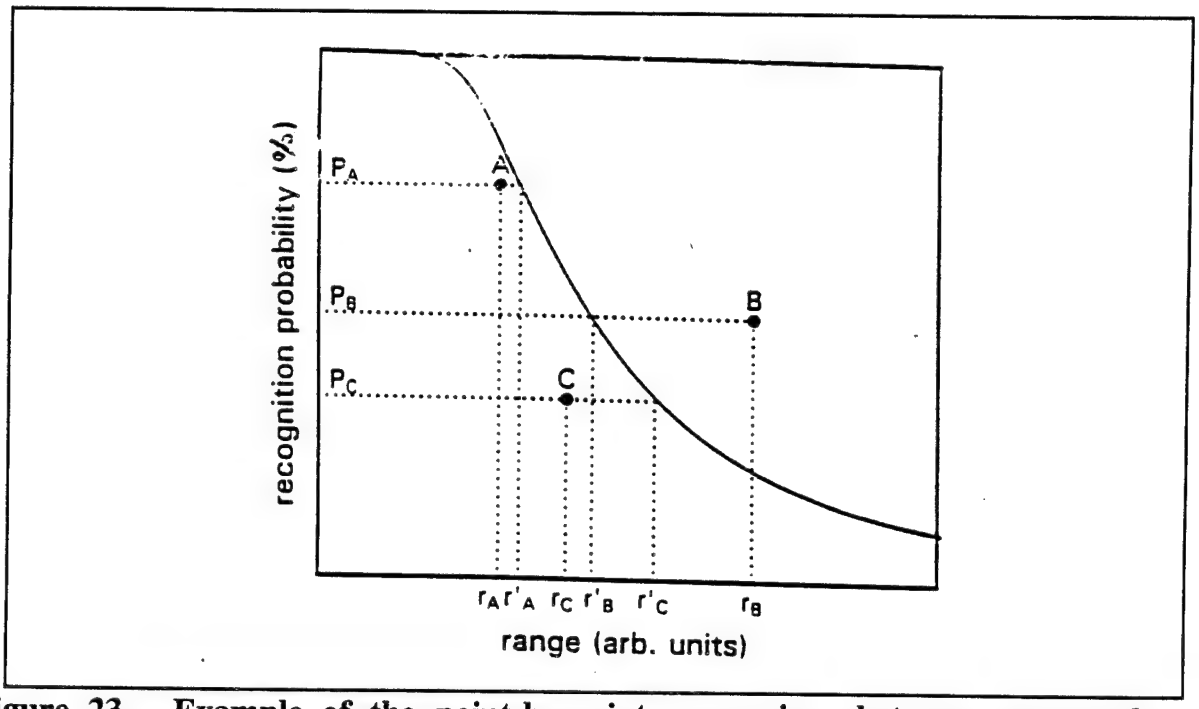


Figure 23. Example of the point-by-point comparison between measured and predicted recognition performance: model prediction (solid line). A, B, and C are datapoints. For each datapoint, probability *P* corresponds to an actual target range *r* and a predicted target range *r'*. For point A, the predicted and measured range are almost identical: ratio $r/r' \approx 1$. For point B, the actual range is longer than the predicted range at the same probability level: $r/r' \approx 1.5$. For point C, the actual range is much smaller than the predicted range at the same probability level: $r/r' \approx 0.75$.

A set of datapoints gives a dimensionless distribution of log (r/r') - values. An example of a distribution is given in figure 24. Mean and variance of the distribution directly provide a measure of the accuracy of the model. The mean of the distribution, «log (r/r')», indicates how well the model predicts overall mean performance. If «log (r/r')» = 0, overall mean performance is correctly predicted. A shift of the distribution along the log (r/r') axis means that predicted acquisition range is too long or too short on average. Therefore, «log (r/r')» provides a range correction factor that will make the model predict overall mean performance correctly.

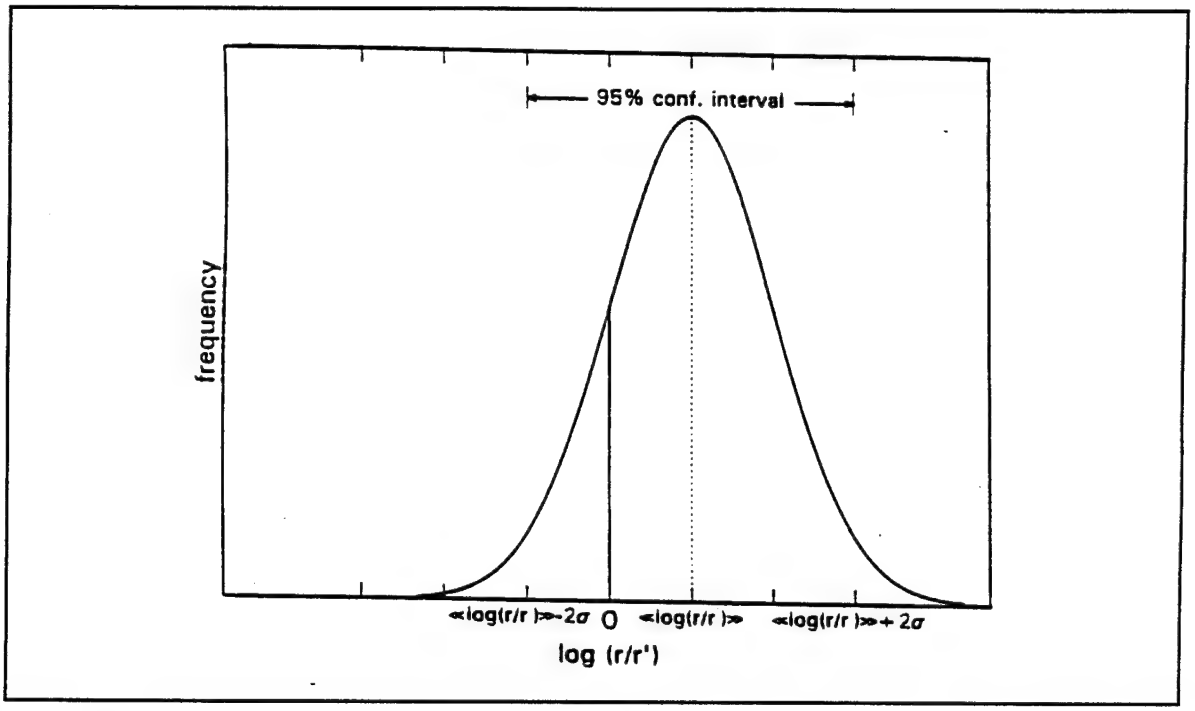


Figure 24. Example of a distribution of r/r' values on a logscale. If the mean of the distribution «log (r/r')» is equal to 0, the model correctly predicts overall mean performance. A shift of the distribution means predicted acquisition range is too long or too short on average. (In the example, predicted ranges are too short.) The variance σ^2 in the distribution indicates how well the model predicts acquisition performance for individual trials.

The variance $\sigma^2_{(r/r')}$ in the distribution indicates how well the model predicts acquisition performance for individual trials. Part of the variance, $\sigma^2_{(r/r'),obs}$, is due to statistical errors in the observer scores because an error in the recognition probability P leads to an error in r', and hence in log (r/r'). Appendix C shows how $\sigma^2_{(r/r'),obs}$ is calculated from the standard error σ_P in P. The remainder of the variance can be ascribed to incorrect predictions by the model. Thus, if the evaluation yields that $\sigma^2_{(r/r')} \approx \sigma^2_{(r/r'),obs}$, the model will be accepted because it correctly predicts observer performance within the accuracy of the measurements, as a function of the parameters that were varied in the experiment (target type, POD).

On the other hand, if $\sigma^2_{(r/r')} \gg \sigma^2_{(r/r'),obs}$, the unexplained variance is mainly due to differences between the model predictions and actual observer performance.

Therefore, the 95 percent confidence interval of the distribution, [«log (r/r')» - $2\sigma_{(r/r')}$, «log (r/r')» + $2\sigma_{(r/r')}$], indicates the quality of the model predictions. A wide uncertainty interval means that the model is not able to make reliable predictions for individual trials.

An elegant property of the procedure is that the entire set of datapoints, irrespective of the run or the circumstances under which they were collected, can be analyzed at the same time, yielding a single distribution. The analyses can also be carried out for subsets of the data. Section 6.7 shows that analyses of subsets may be used to discover which factors significantly contribute to the variance in the distribution. The model may be improved if these factors are known.

6.6.3.2 *Results.*— Figure 25 shows the results of the evaluation for data set A (38 runs, pop-up presentation order). Figure 25a plots the ratio between actual and predicted range (r/r') for individual datapoints. Only datapoints with a recognition probability between 20 and 80 percent are considered (230 points), because r/r' may take unrealistic values (section 6.6.3.3) at very high or very low probabilities. The standard deviation $\sigma_{(r/r'),obs}$ in the (r/r') values, caused by the statistical error in the observer scores, is presented in the upper right-hand corner of the figure. Appendix C shows that $\sigma_{(r/r'),obs}$ = 0.03 - 0.06 log units for probabilities between 20 and 80 percent. Figure 25b shows the histogram of the distribution of log (r/r')-values, based on the datapoints in figure 25a. Note that this is a roughly normal distribution. The mean and the 95 percent confidence interval of the distribution are indicated in figure 25a by the fat dashed line and the two dotted lines, respectively.

It is clear, that the quality of the model predictions for individual trials is not very good. First, the mean of the distribution is 0.23, which means that actual ranges are $10^{0.23}$ = 1.70 times the predicted ranges on average. The average range correction factor is 1.8 for the whole study. This difference cannot be ascribed to inaccuracies in the values of the input parameters (see 6.4.5).

Second, the standard deviation $\sigma_{(r/r')}$ = 0.17 on a log scale, which is much larger than $\sigma_{(r/r'),obs}$. The 95 percent confidence interval is [-0.11, 0.57]. This

interval is [0.8, 3.7] on a linear scale, spanning a factor of almost 5. There is a 95 percent probability that the actual recognition range for an individual trial falls between 0.8 and 3.7 times the range that is predicted by TARGAC. Even after correction for the overall mean acquisition range, the actual range may be more than twice or less than half the predicted range. Thus, the model is not very good at predicting how observer performance depends on the prevailing conditions in the field.

For data sets B (15 runs, approaching presentation order) and C (15 runs, pop-up presentation order), similar results are found. The mean shifts between the data and the TARGAC predictions are 0.28 for set B (70 datapoints) and 0.27 for set C (65 datapoints) on a log scale, which correspond to an underestimate of the actual range by a factor 1.90. The uncertainty interval is very large in both cases, spanning a factor of about 4 for the pop-up presentation order (data set C) and 3.3 for the approaching presentation order (data set B). The smaller interval for the latter case is found because the effects of target/terrain interactions are less pronounced using this presentation order.

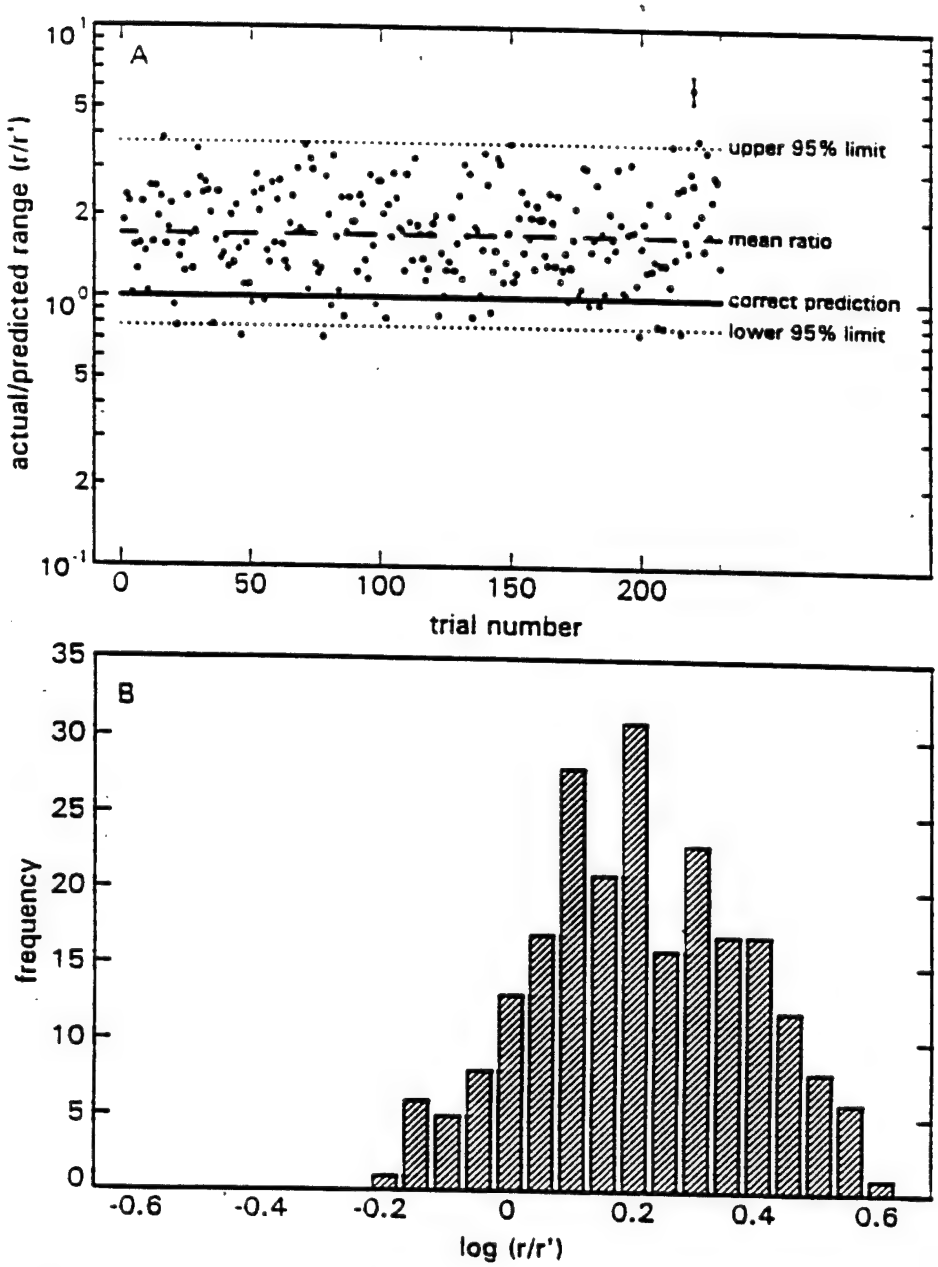


Figure 25. Comparison between measured and predicted recognition performance for individual trials: (a) ratio between actual and predicted range (r/r'); (with a perfect model, the points would be concentrated around the solid line ($r/r' = 1$). The dashed line corresponds to the mean of the log (r/r') distribution. The dotted lines indicate the boundaries of the 95-percent confidence interval. The error in observer scores is shown in the upper right-hand corner of the figure.) (b) histogram of the log (r/r') distribution (mean 0.23; standard deviation 0.17). It is clear that the model is not very good at predicting recognition performance for individual trials.

6.6.3.3 *High and Low Probability Levels.*— At very high or low probabilities, small errors in the observer scores lead to large errors in the value of r/r', because the slope of the predicted probability versus range curve is very shallow in those regions. Appendix C shows that for the analyses it is safe to use only observer data with a probability level between 20 and 80 percent. Using data with a higher or lower probability level may lead to a broadening of the log (r/r') distribution. The effect of using a larger interval on the distribution was estimated for the three data sets A, B, and C.

The effect was similar for all sets. Both mean and standard deviation of the log (r/r')-distribution are only slightly increased (approximately by 0.02 and 0.01 log units, respectively) when the interval is changed from 20 to 80 percent to 10 to 90 percent. A further increase is not possible because TARGAC predictions are only defined between 10 and 90 percent. As a conclusion, the extent of the interval does not have a very large effect on the results.

6.6.4 Conclusions

There are important differences between recognition performance, as measured in observer experiments, and the TARGAC predictions for the BEST TWO situation:

1. The model predictions are too conservative. TARGAC underestimates recognition range by a factor 1.8 on average. This ratio is similar for pop-up and approaching targets. The difference cannot be ascribed to inaccuracies in the values of the model input parameters.

2. TARGAC does not make accurate predictions for individual trials. The analyses show that the 95 percent-confidence interval is roughly given by 0.9 to 3.6 times the predicted acquisition range, spanning a range of a factor of 4.

6.7 Possible Sources of the Unexplained Variance

The previous section shows that there is a large amount of unexplained variance when TARGAC predictions are compared with actual observer performance for individual trials. The variance cannot be ascribed to the

statistical error in the observer scores. The conclusion is that the model does not predict how observer performance depends on field factors or parameters that were varied in the experiment.

It may be possible to make a model that better predicts performance for individual situations if the factors that contribute significantly to the unexplained variance can be determined. Such factors can be found using Analysis of Variance. Suppose that the effect of target type is not modeled correctly, the broad distribution of log (r/r') values that were found for the entire dataset is in fact a composition of narrower distributions with different mean shifts for different target types. The model could be improved (the amount of unexplained variance would be reduced) by remodeling the effect of target type.

In a similar way, possible effects of other factors may be tested. These hypotheses can be straightforwardly tested with the present data set:

a. Target type has an effect on acquisition range, but the effect is not modeled adequately by simply taking its minimum dimension.

b. Time of day has an effect on observer performance, although the model does not predict any differences.

c. Part of the variance that was found in data set A, may be due to using stationary and (head-on) moving targets. The model is only designed for stationary targets.

d. Target-terrain interaction, section 6.6.1, has a considerable impact on acquisition performance. This hypothesis can be verified by comparing the results for different approach routes.

Analysis of Variance was used to determine which of the above-mentioned factors have a significant effect on the log (r/r') distribution for the comparison between the TARGAC predictions and the observer data from set A, which is the largest data set. Only main effects are considered; interactions are not considered relevant for a first investigation.

The analyses show that there are statistically significant effects ($P < 0.05$) of target type, time of day, and approach route on variance. No significant effect was found for head-on target motion. The effects follow:

6.7.1 Target Type

Mean range shifts («log (r/r')») for most targets are quite similar to the mean shift found for the complete set (ranges do not differ by more than 20 percent). However, for one target (AMX-10) the shift is considerably larger. Predicted range for this target is much shorter than for the other targets because of its small height (see table 12), but measured performance is slightly better than for the other targets. The standard deviation of the distribution is reduced from $\sigma = 0.17$ to $\sigma = 0.15$ when optimal shifts are applied for each target type separately, which means that most of the variance remains unexplained. Thus, the model cannot be improved considerably by remodeling the effect of target type.

6.7.2 Time of Day

Runs were divided into three categories: morning, afternoon, and early night in accordance with the division that was made during the BEST TWO trials. [19,20] There is a small but significant difference in mean range shift for morning and afternoon and early night. There is no relevant reduction of the width of the distribution when the effect of time of day is taken into account.

6.7.3 Approach Route; Target/Terrain Interactions

Mean recognition ranges for the right approach route were approximately 20 percent longer than for the left route. The routes were very close to each other, heading in approximately the same direction. No theoretical explanation exists for acquisition of targets on the right approach route being easier. This means that local factors are very important. Furthermore, the results of the observer experiments also show that for a single approach route, in general, there is no monotonous relationship between recognition performance and target distance. The effects were ascribed to a strong interaction between

target signature and local background (target/terrain interactions, section 4). A large amount of variance may be ascribed to target/terrain interactions.

A model that takes into account the effects of target/terrain interaction may become very complicated. Apart from mean or area contrast between target and background, which is modeled in TARGAC, there are many local factors that may influence acquisition performance, such as edge contrast, internal target contrast, differences in target and background structure, or variations in target orientation. Their effect on acquisition performance is unknown. Analyzing the BEST TWO images, may make it possible to determine the effect of some of the above mentioned local factors on recognition performance. Such analyses go beyond the scope of this report.

In conclusion, there seems to be no simple modification that would lead to a model that better predicts acquisition performance for individual trials. (A modification that would considerably diminish the amount of unexplained variance that is found when TARGAC predictions are compared with actual observer performance.)

6.8 Discussion

TARGAC is a very comprehensive TA model. The model combines modules for target and background characterization, atmospheric transmittance, and system/observer performance. These properties make TARGAC a very useful tool for military purposes, especially as a TDA. However, the evaluation of TARGAC shows that the model does not predict recognition performance very accurately. The main differences between observed recognition performance and the model predictions for the BEST TWO situation follow:

1. TARGAC underestimates the mean recognition range for the BEST TWO situation by a factor of about 1.8.

2. TARGAC predictions for individual cases have an uncertainty interval of roughly 0.9 to 3.6 times the predicted acquisition range, spanning a range of a factor of 4.

The sensitivity analyses (section 4) showed, that for the BEST TWO situation only the module that describes EO and human visual system performance, is responsible for the range predictions. The effects of the outcome of the target background contrast module and the atmospheric transmittance module on the range predictions are negligible because of the excellent conditions. The system performance module in TARGAC is theoretically equivalent to 1-D ACQUIRE90, the 1-D option of the 1990 version of the NVESD Static Performance Model (the 2-D option of ACQUIRE is discussed below). Figure 26 presents a comparison between the TARGAC and the 1-D ACQUIRE90 recognition performance predictions for the BEST TWO standard situation (section 6.4.2). Filled circles represent the TARGAC predictions, and open circles represent the corresponding 1-D ACQUIRE90 predictions. Atmospheric transmission was set to 1.0, and target background contrast was set to 5.0 K in ACQUIRE90. Varying the contrast had little effect on the ACQUIRE90 predictions. Evidently, the predictions made with the two models are identical.

The conclusion is that the results of the evaluation are not only relevant for TARGAC but for all models that are based on the 1-D NVESD Static Performance Model.

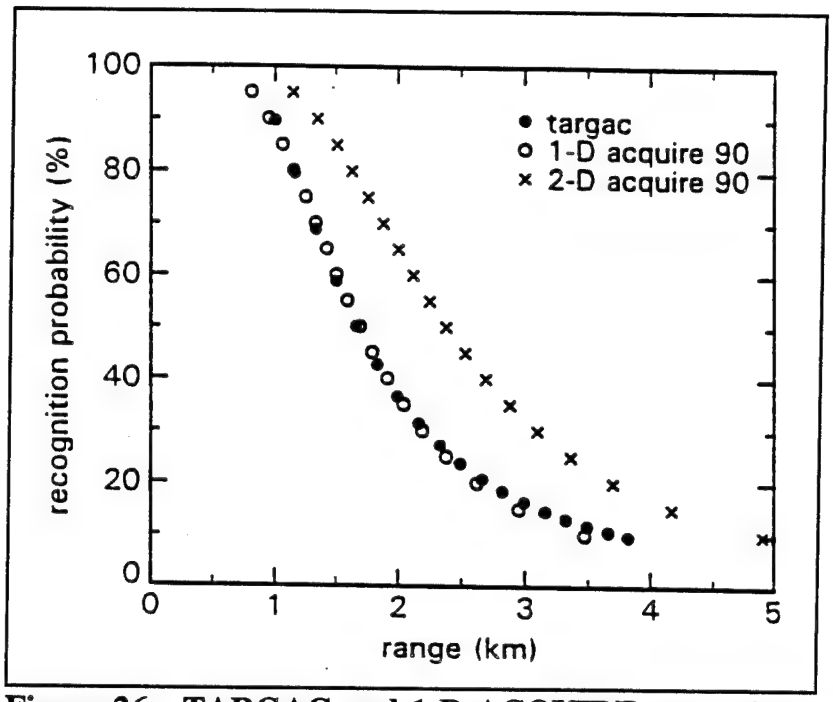


Figure 26. TARGAC and 1-D ACQUIRE recognition performance predictions for the BEST TWO situations are identical. 2-D ACQUIRE predicts longer ranges.

6.8.1 Mean Acquisition Performance

Section 6.4 showed that inaccuracies in the data input to the model may lead to an error of up to 20 percent in predicted range. The performance of the observers in the experiment may be relatively high, because they were trained on six targets in a specific situation (a single weather condition, one terrain). Their score might have been lower if there were more uncertainties in their task. However, the factors are not large enough to explain the considerable difference between actual and predicted mean acquisition performance.

TARGAC may easily be adapted to correctly predict mean acquisition performance for the BEST TWO situation by changing the recognition criteria in the system performance module. The present version of the model predicts a recognition probability of 50 percent if four line pairs can be resolved across the effective dimension of the target. According to Ratches, [13] this is a conservative criterion (section 6.2.1.3). However, even the optimistic criterion that Ratches gives (a three line pair criterion, which would predict 4/3 longer

ranges) is still too conservative. A recognition probability of 50 percent should correspond to the resolution of 2.2 line pairs across the effective dimension of the target to correctly predict mean recognition range for the BEST TWO situation. It should also be noted that the targets in BEST TWO were always in front view. It is well known that targets in side view are more easily recognized than targets in front view, but the 1-D Static Performance Model predicts equal ranges because the effective dimension is target height in both cases. This means that, for targets in side view, prediction of mean recognition performance by TARGAC may be even further off.

Recently, a new version of the Static Performance Model (2-D ACQUIRE90) was developed; it makes its predictions on the basis of two target dimensions and the horizontal and vertical MRTD. Figure 26 shows that the new version (crosses) predicts longer ranges than the 1-D model. It also predicts better acquisition performance for targets in side view than for targets in front view. For the BEST TWO situation, the 2-D version underestimates mean acquisition range by less than 30 percent, an error which is near the accuracy of the MRTD cut-off frequency. TARGAC may be improved by implementing the 2-D version of the Static Performance Model.

6.8.2 Acquisition Performance for Individual Cases

The TARGAC user interface suggests that predictions can be made for individual targets under given conditions. For example, the model predicts the 50 percent recognition range for a T62 tank on a grass field at 1400. The evaluation shows that the model cannot make such predictions. A very large uncertainty interval is found when the predictions are compared with data from a single observer performance experiment, using one target set, in one terrain, with one weather condition, and one camera. The interval may be even larger if more conditions are investigated. The large amount of unexplained variance is not due to possible errors in the input data, such as the MRTD-curve, because these only affect the mean of the confidence interval and not its width. The width of the interval is not decreased if the original system performance model is replaced by the 2-D version. Apparently, the model is not sophisticated enough to deal with individual cases: factors that play an important role in TA are not modeled or are not modeled correctly.

A model that better predicts acquisition performance for individual trials may have to be very complex. Section 6.7 showed that local factors play an important role in TA. A correct prediction probably requires a detailed model of the effects of local factors on acquisition performance and a detailed description of the local conditions as input to the model. Such a detailed modeling, if possible at all, may not be useful for practical purposes. The equivalent disc model [21] is based on a different approach; it predicts acquisition performance for a set of targets, rather than predicting a range for each target separately.

It is likely that the width of the uncertainty interval depends on the type of terrain. For the BEST TWO trials, local factors were very important because of the inhomogeneity of the background and, consequently, the uncertainty interval is large. If the background is more uniform, one expects that acquisition performance mainly depends on target distance (as a model predicts), resulting in a smaller 95 percent confidence interval. The dependence of the reliability of the acquisition range predictions on terrain type or on statistical information about the terrain, may be a topic of future research.

6.9 Conclusions and Recommendations

The TA model TARGAC was evaluated using BEST TWO observer performance data for recognition of targets in front view. Because of the excellent atmospheric conditions during BEST TWO, recognition performance predictions are determined solely by the system performance module of TARGAC, which is equivalent to the 1-D NVESD Static Performance Model. The main results of the evaluation follow:

1. The model predictions are too conservative. TARGAC underestimates recognition range by a factor 1.8 on average. This ratio is similar for pop-up and approaching targets.

2. TARGAC does not make accurate predictions for individual cases. The analyses show that the uncertainty interval roughly ranges from 0.9 to 3.6 times the predicted acquisition range, spanning a range of a factor of 4.

3. The TARGAC predictions for overall mean performance can be improved by incorporating the 2-D version of the Static Performance Model.

4. It is proposed that TARGAC predictions are not only presented as single numbers for acquisition probability versus target range, but that some indication is given of the accuracy of the results, preferably in the form of a 95 percent confidence interval.

5. The version of TARGAC tested (PC version released in 1992) contained a number of software errors and minor problems. A number of corrections are suggested. Additional work in modularizing and streamlining the model is recommended. It is also recommended that the model is given a more consistent and user-friendly interface.

6. TARGAC and other models that incorporate the NVESD Static Performance Model should only be used to provide an indication of the actual acquisition performance.

References

1. Valeton, J. M., P. Bijl, and A. J. C. De Reus, *The BEST TWO Field Trial: Timetables of Events, Maps, and Target Distance Data*, Rep. No. IZF 1992 A-10, TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1992.

2. Danielian, F., "An Overview of the RSG 15 Test," In *Proceedings of the Second Symposium on Measuring and Modeling the Battlefield Environment*, Paris, NATO, pp. 1, 1992.

3. Vonhof, D. M., and A. Goessen, "The Effect of White Phosphorous Smoke and Artillery Dust Clouds on Target Acquisition," In *Proceedings of the Second Symposium on Measuring and Modeling the Battlefield Environment*, Paris, NATO, pp. 19, 1992.

4. Lt. Col. D. M. Vonhof, OCI, Royal Netherlands Army, personal communication.

5. Smith, R., and T. Corbin, "Meteorological Conditions and Data for the BEST TWO Field Trial," In *Proceedings of the Second Symposium on Measuring and Modeling the Battlefield Environment*, Paris, NATO, pp. 3, 1992.

6. Wagenaars, W. M., "Localization of Sound in a Room with Reflecting Walls," *J. Audio Eng. Soc.* 38(3), 99-110, 1990.

7. Clement, D., personal communication, Principal to NATO AC243/RSG.15, Forschungsinstitut fur Optik, Schloss Kressbach, D-72072, Tuebingen, Germany.

8. Vonhof, D. M., and J. Rogge, "Reliability of Observer Responses Using Inventory Thermal Imagers," In *Proceedings of the Second Symposium on Measuring and Modeling the Battlefield Environment*, Paris, NATO, pp 23, 1992.

9. Wester, H. G., and F. W. M. Van de Mortel, *Waarnemings Experiment BEST TWO* (in Dutch), Report, Royal Military Academy, Breda, The Netherlands, 1990.

10. Bartleson, C. J., and F. Grum, "Optical Radiation Measurements," Vol. 5: *Visual Measurement*, Academic Press, 1984.

11. Sanders, J. S., M. S. Currin, and C. E. Halford, "Visual Perception of Infrared Imagery," *Optical Engineering*, **30**, p 1674-1681, 1991.

12. Johnson, J., "Analysis of Image Forming Systems," In *Image Intensifier Symposium*, Warfare Vision Branch, Electrical Engineering Dept., U.S. Army Engineer Research and Development Laboratories, Fort Belvoir, VA, pp. 249-273, 1958.

13. Gillespie, P., *TARGAC User's Guide*, U.S. Army Research Laboratory, Battlefield Effects Directorate, White Sands Missile Range, NM, USA, 1990.

14. Ratches, J. A., W. R. Lawson, F. J. Shields, C. W. Hoover, L. P. Obert, S. P. Rodak, and M. C. Sola, *Status of Sensor Performance Modeling at NV&EOL*, Night Vision & Electro-Optics Laboratory, Fort Belvoir, VA 22060, USA, 1981.

15. Ratches, J. A., "Static Performance Model for Thermal Imaging Systems," *Opt. Eng.* 15, 6, pp. 525-530, 1976.

16. Jong, A. N. de, Y. H. L Janssen, M. J. J. Roos, and R. A. W. Kemp, *Infrared and Electro-Optical Experiments During Best Two by Research Group Infrared*, Report No. FEL-91-A252, TNO Physics and Electronics Laboratory, The Hague, The Netherlands (NATO Restricted), 1991.

17. Andersen, E., *BEST TWO Infrared Signatures of the Vehicles Participating in the Trial*, Report No. DDRE N-7/1991, Danish Defence Research Est. Copenhagen, Denmark (NATO Confidential), 1991.

18. Bakker, S. J. M., and M. J. J. Roos, *MRTD-metingen aan warmtebeeldcamera's* (in Dutch), Rep. No. FEL-1989-118, TNO Physics and Electronics Laboratory, The Hague, The Netherlands (NATO Confidential), 1989.

19. Valeton, J. M., and J. Rogge, *The BEST TWO Scenarios: An Overview*, Rep. No. IZF 1992 A-32, TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1992.

20. Reichart A. (Ed.), *Second Symposium on Measuring and Modeling the Battlefield Environment*, Volume 1A (unclassified). Technical Proceedings AC/243 (Panel 4) TP/1. SEFT/CTE/OSA, 1993.

21. van Meeteren, A., "Characterization of Task Performance with Viewing Instruments," *J. Opt. Soc. Am.* 7, 10, pp. 2016-2023, 1990.

22. Gillespie, P., personal communication, ARL, White Sands Missile Range, NM, 1993.

Acronyms and Abbreviations

AAODL	Atmospheric Aerosol and Optics Data Library
ACQUIRE	target acquisition model developed by NVESD
APC	armored personnel carrier
ARL	Army Research Laboratory
AUTOFEDS	system developed by NVESD for recording target vehicle movements and observer responses during field trials. No longer used.
BED	Battlefield Environment Directorate
BEST TWO	Battlefield Emissive Sources Trials under the European Theater Weather and Obscurants
CCD	charge-coupled device
D	detection
EO	electro-optical
EOSAEL	Electro-Optical Systems Atmospheric Effects Library
ETCA	Etablisement Centrale Technique d'Armement
FLIR	forward looking infrared radar
I	identification
LMT	French battlefield radar system
LUST	limited use smoke technology device
MIA	main instrumentation area
MRC	minimum resolvable contrast
MRTD	minimum resolvable temperature difference
NATO	North Atlantic Treaty Organization

NVESD	Night Vision Electro-Optics Sensors Directorate
1-D	one-dimensional
POD	part of day
R	recognition
RASIT	French battlefield radar system
TA	target acquisition
TARGAC	Target Acquisition Model
TCM2	Thermal Contrast Model
TDA	tactical decision aid
TNO	Netherlands Organization for Applied Scientific Research
TNO-FEL	TNO Physics and Electronics Laboratory
TNO-HFRI	TNO Human Factors Research Institute
TOW	tube launched optically tracked wire-guided missile
TTPF	target transfer probability function
2-D	two dimensional
VIS	visual
ZIL	a Soviet truck

Bibliography

Bijl, P., and J. M.Valeton, *Observer Experiments with BEST TWO Thermal Images. Part 2: Terrain Interaction and Target Motion*, Rep. No. IZF 1992 A-34, TNO Human Factors Research Institute, Soesterberg, The Netherlands, (NATO Confidential), 1992.

Bijl, P., and J. M. Valeton, *Observer Experiments with BEST TWO Thermal Images. Part 3: Reliability of Observer Responses*, Rep. No. IZF 1992 A-35, TNO Human Factors Research Institute, Soesterberg, The Netherlands, (NATO Confidential), 1992.

Bijl, P., and J. M. Valeton, *Evaluation of Target Acquisition Models TARGAC Using BEST TWO Observer Performance Data*, Rep. No. IZF 1994 A-??, TNO Human Factors Research Institute, Soesterberg, The Netherlands, (In press), 1994.

Valeton, J. M., and P. Bijl, *Observer Experiments with BEST TWO Thermal Images. Part 1: Design, Training and Observer Selection*, Rep. No. IZF 1992 A-33, TNO Human Factors Research Institute, Soesterberg, The Netherlands, (NATO Confidential), 1992.

Appendix A

The Complete Set of Observer Performance Data

The complete set of observer response data is presented in 126 plots (figures A-1a-o through A-8a-o). All the results are presented as the percentage of correct identification and recognition responses versus target range. The standard deviation of the datapoints is typically about 10 to 20 percent.

Figures A-1 through A-4 (63 plots) represent the data obtained for the forced condition. In this condition (section 5.4), the observers were forced to name the target, even if they were not sure which vehicle was presented. Figures A-5 through A-8 show the data for the unforced condition. The scores are for free (unforced) identification or recognition reports, which correspond better to the Target Acquisition (TA) task in a practical military field situation.

Figures A-1 through A-3 and A-5 through A-7 show the data for the position, or pop-up, presentation order (POS), (section 3.2.4), whereas, figures A-4 and A-8 show the data for the sequential presentation order (indicated as RUN).

The data were obtained in two series of experimental sessions. Acquisition performance was determined for 15 daytime runs (parts of day (POD) 2 and 3) of stationary targets (Scenario 1) in Experiment 1 (figures A-1, A-4, A-5, and A-8). Both types of presentation order (POS and RUN) were applied. The mean scores for 11 observers are presented in the figures. Data was collected for 33 runs of stationary (Scenario 1) and moving (Scenario 2) targets on all part of day (PODs) in Experiment 2 (figures A-2, A-3, A-6, and A-7). Only POS was used. Mean scores are presented for four observers.

A complete overview of the structure of the data is presented in Bijl and Valeton.*

---

*Bijl, P., and J. M. Valeton, *Observer Experiments with BEST TWO Thermal Images. Part 4: Complete Set of Observer Performance Data*, Rep. No. IZF 1992 A-43, TNO Human Factors Research Institute, Soesterberg, The Netherlands, (NATO Confidential), 1992c.

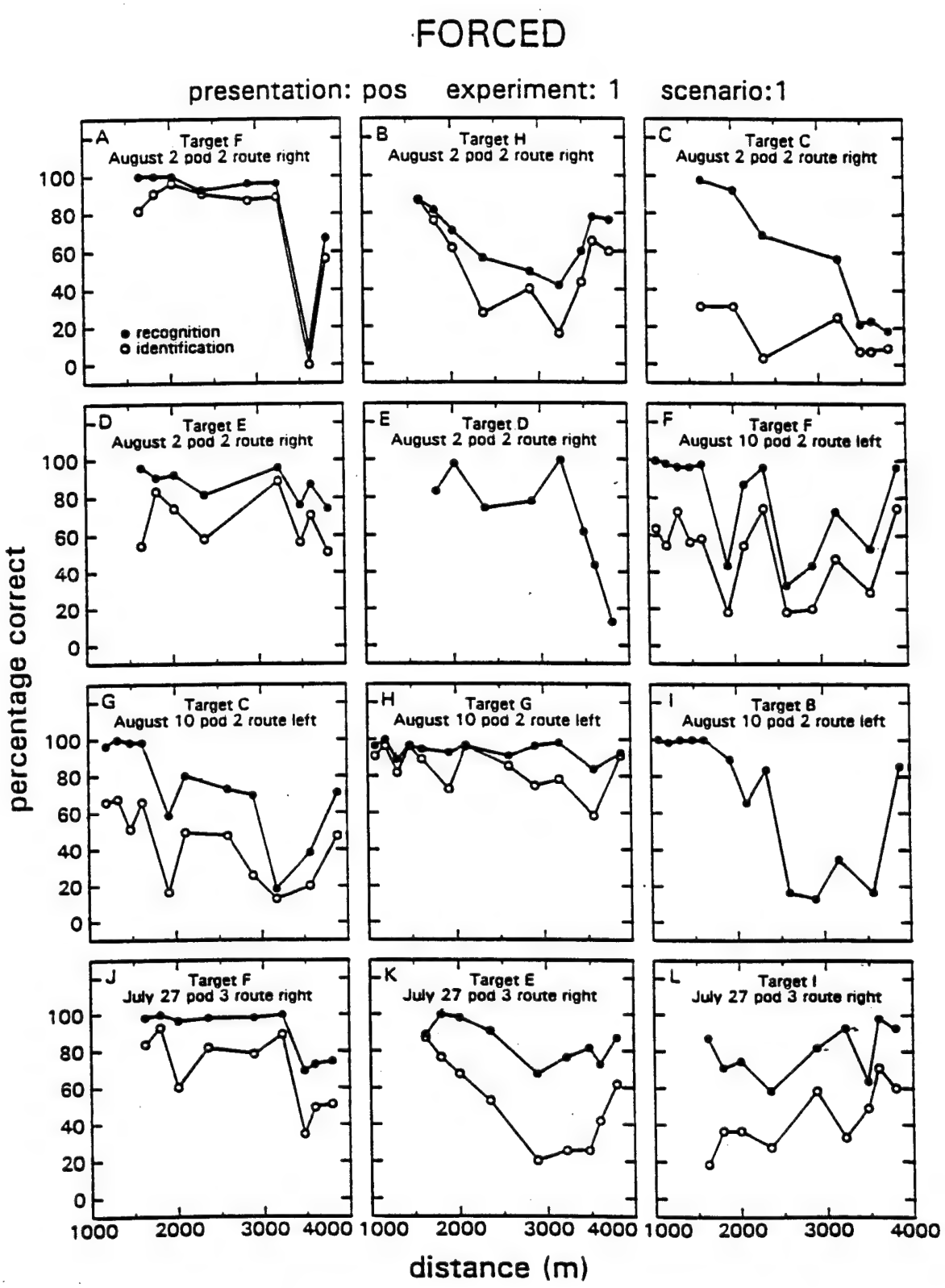


Figure A-1. Observer recognition and identification performance for the condition FORCED-POS-Scenario 1, Experiment 1.

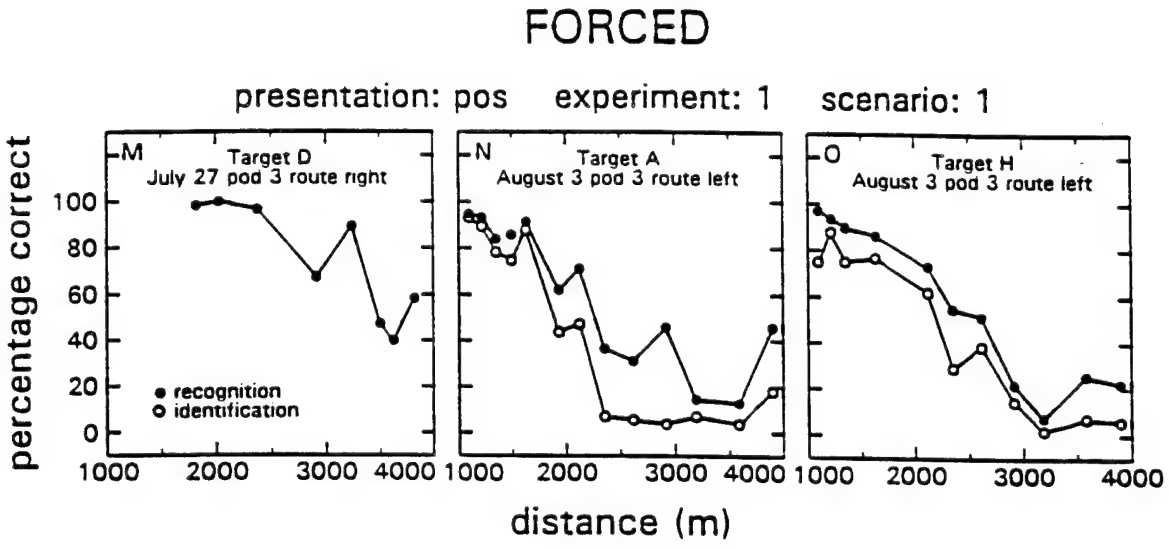


Figure A-1. Observer recognition and identification performance for the condition FORCED-POS-Scenario 1, Experiment 1 (continued).

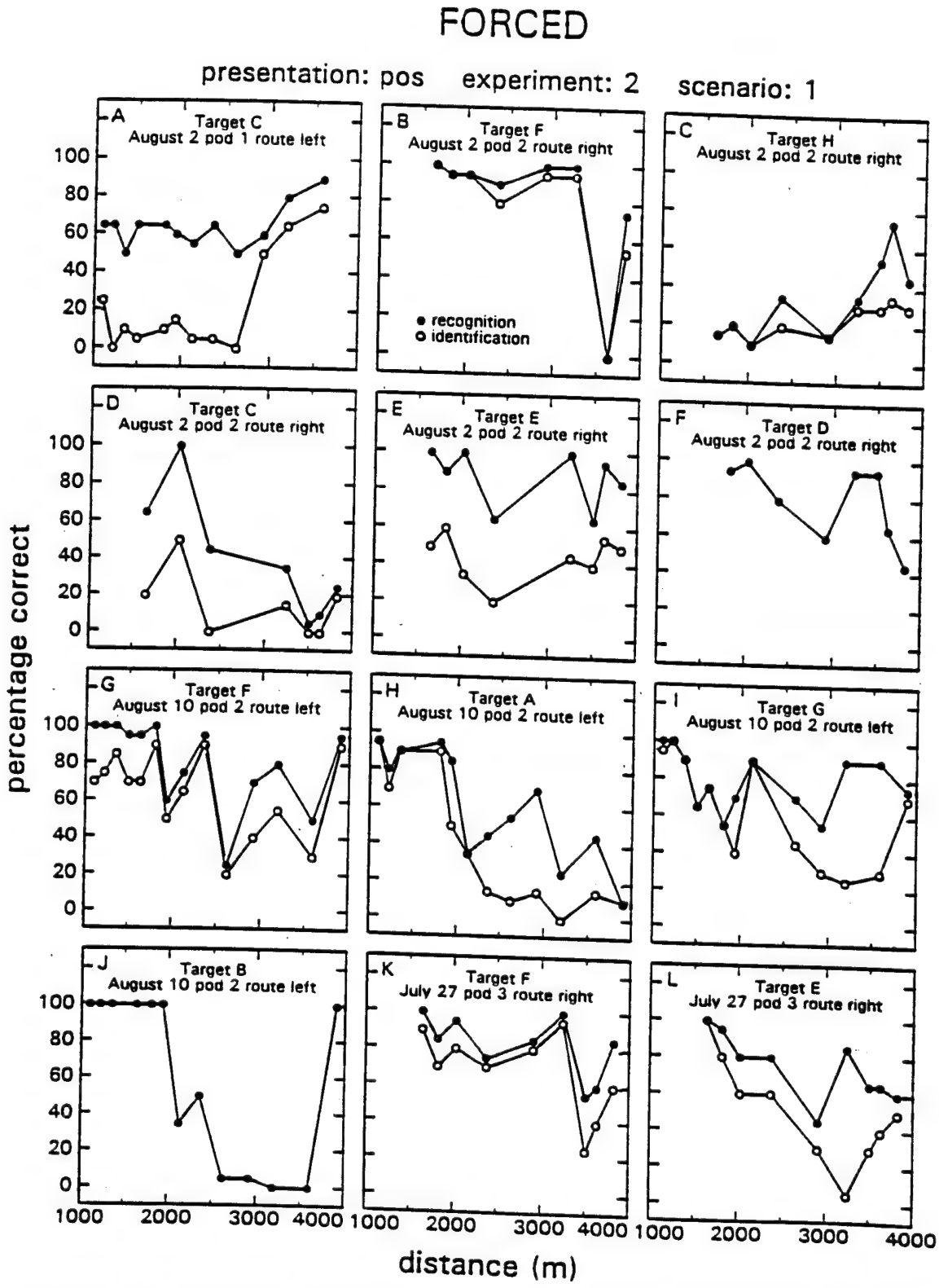


Figure A-2. Observer recognition and identification performance for the condition FORCED-POS-Scenario 1, Experiment 2.

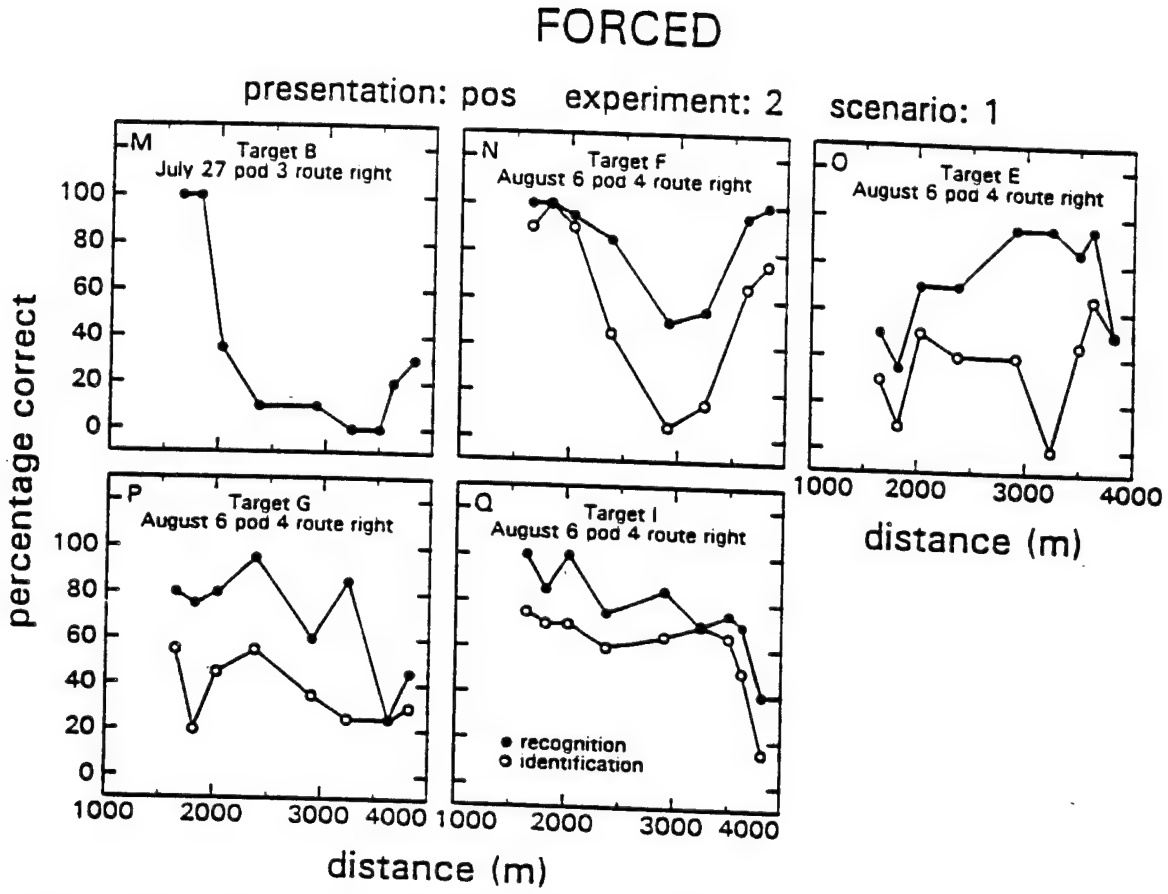


Figure A-2. Observer recognition and identification performance for the condition FORCED-POS-Scenario 1, Experiment 2 (continued).

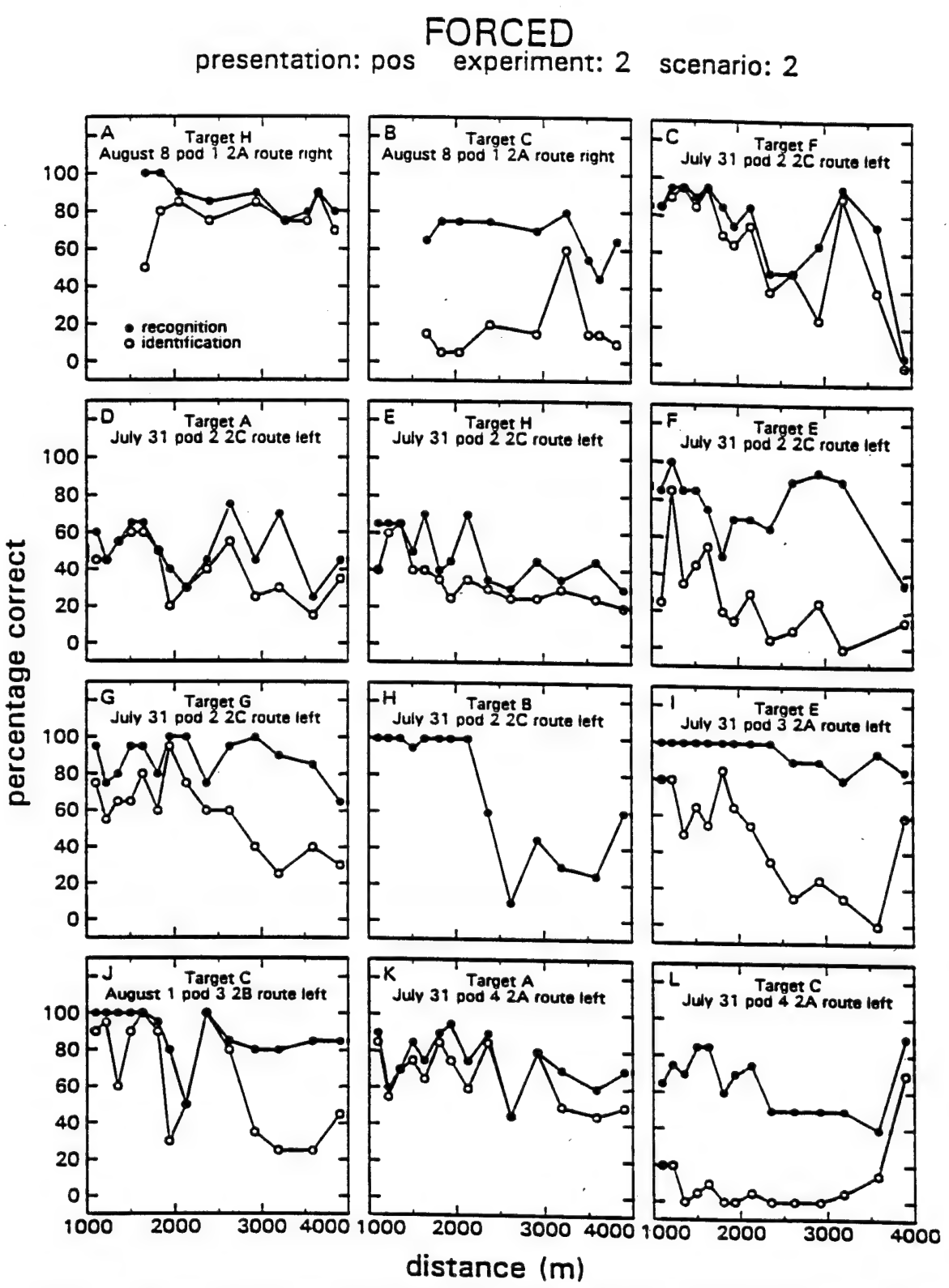


Figure A-3. Observer recognition and identification performance for the condition FORCED-POS-Scenario 2, Experiment 2.

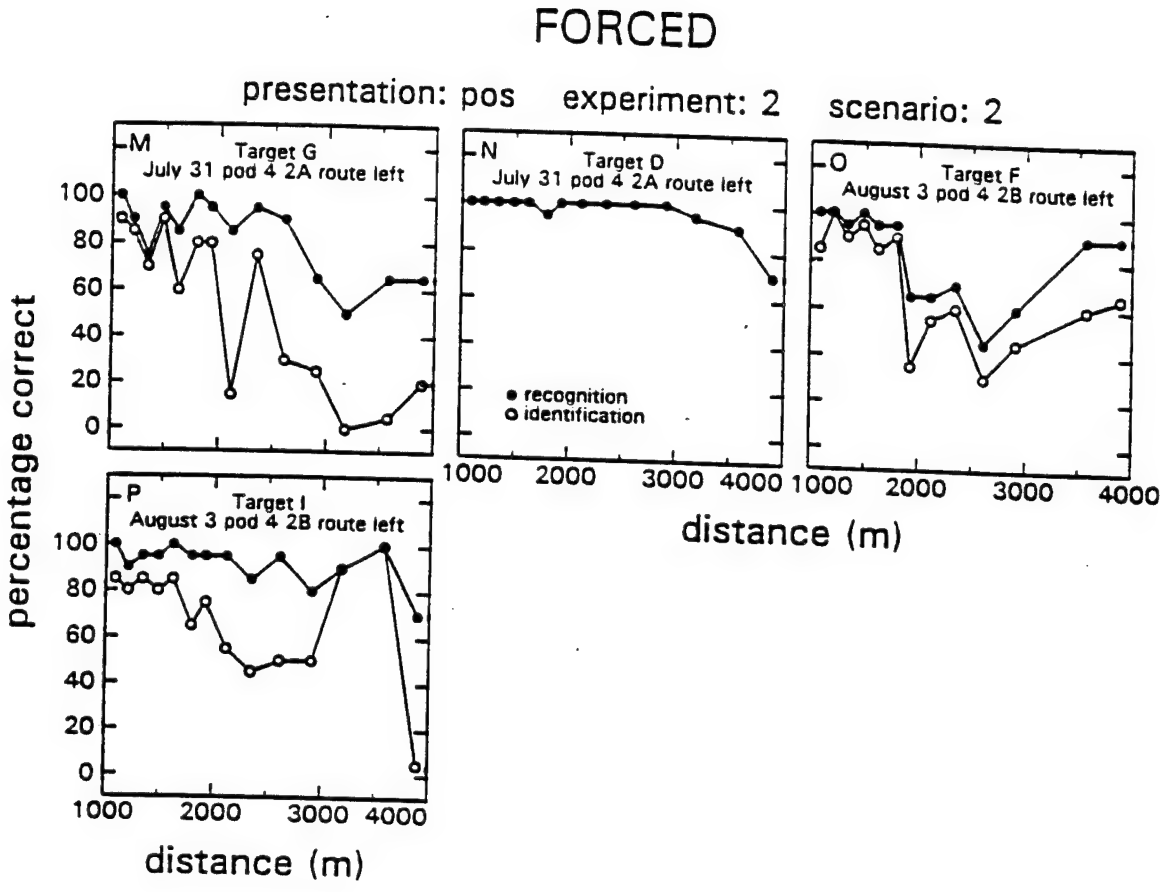


Figure A-3. Observer recognition and identification performance for the condition FORCED-POS-Scenario 2, Experiment 2 (continued).

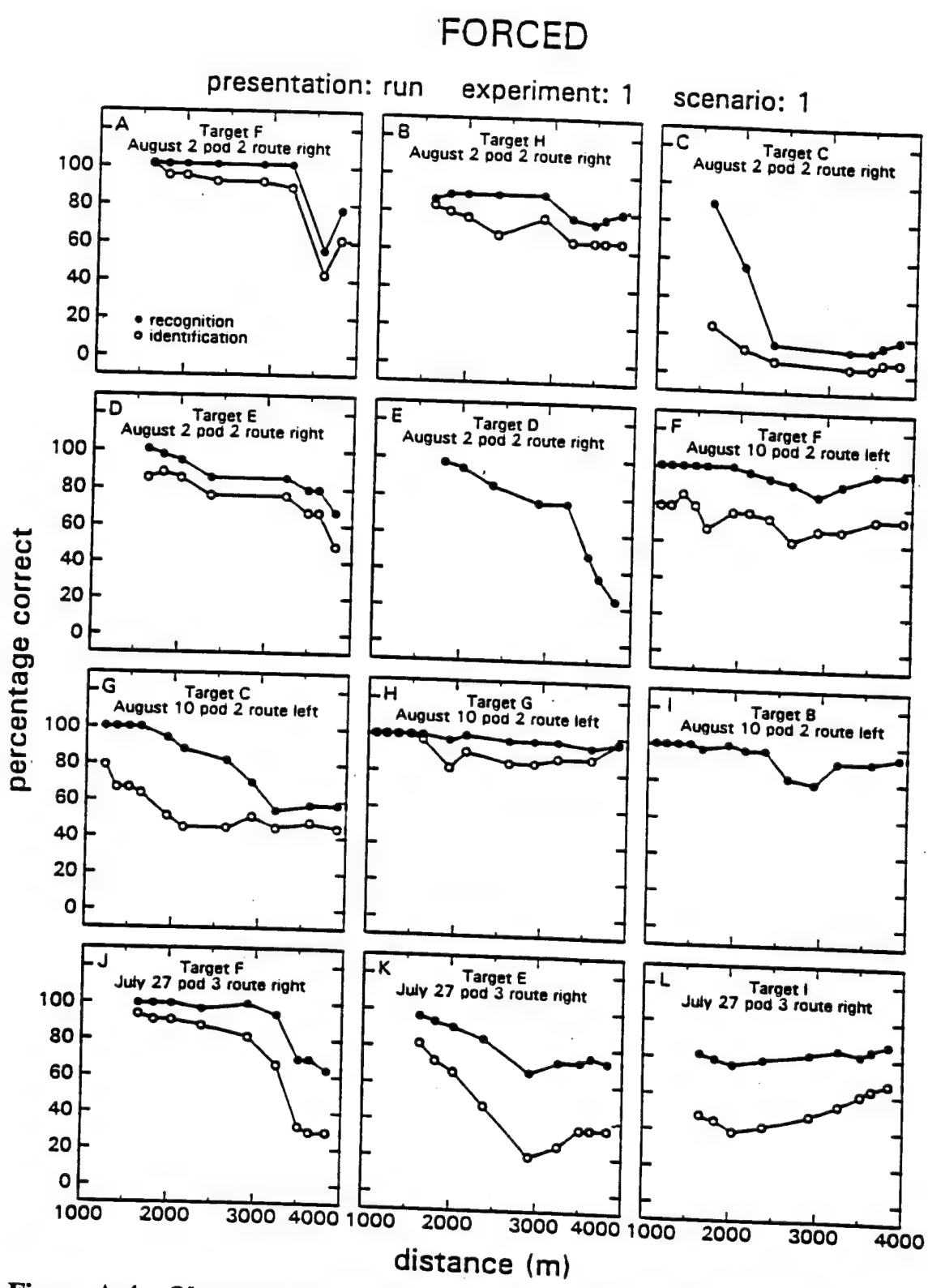


Figure A-4. Observer recognition and identification performance for the condition FORCED-RUN-Scenario 1, Experiment 1.

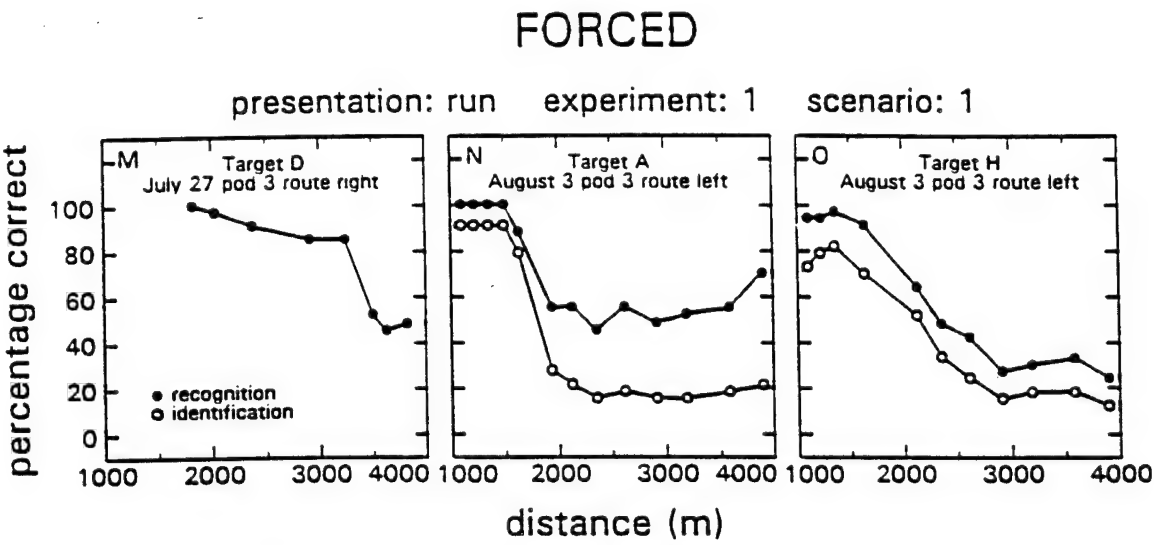


Figure A-4. Observer recognition and identification performance for the condition FORCED-RUN-Scenario 1, Experiment 1 (continued).

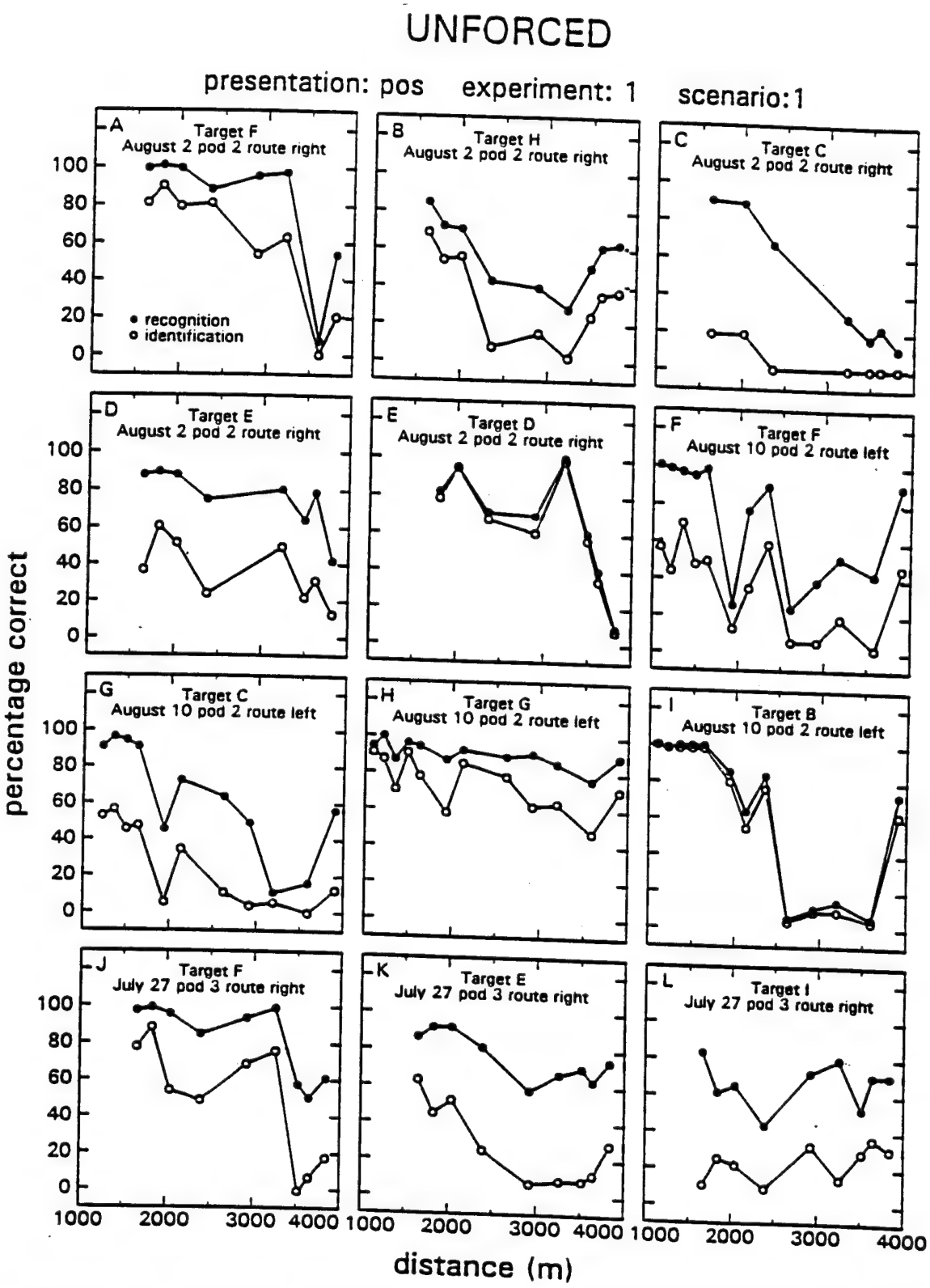


Figure A-5. Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 1, Experiment 1.

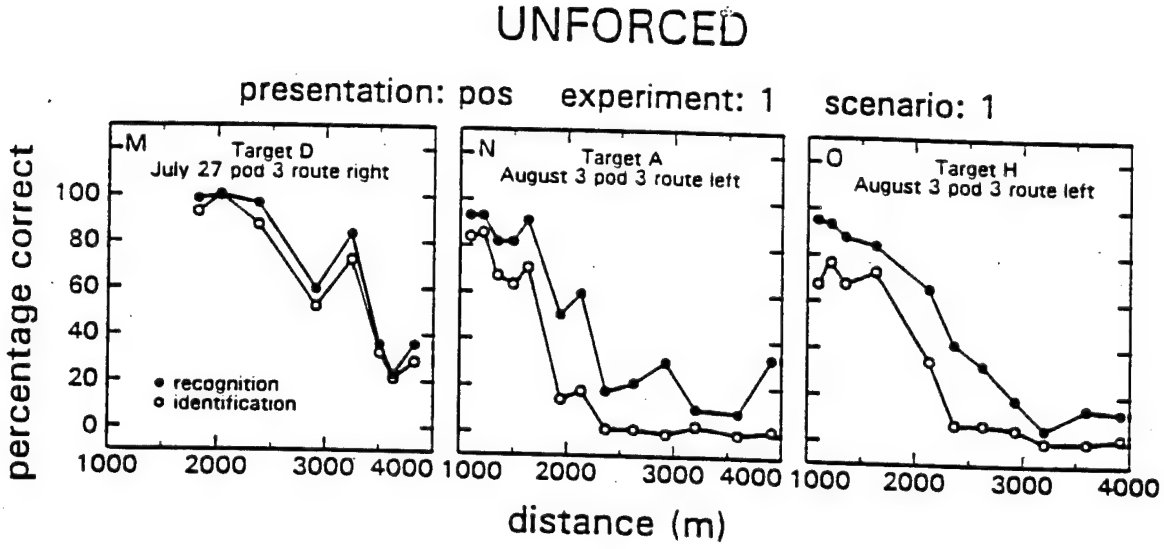


Figure A-5. Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 1, Experiment 1 (continued).

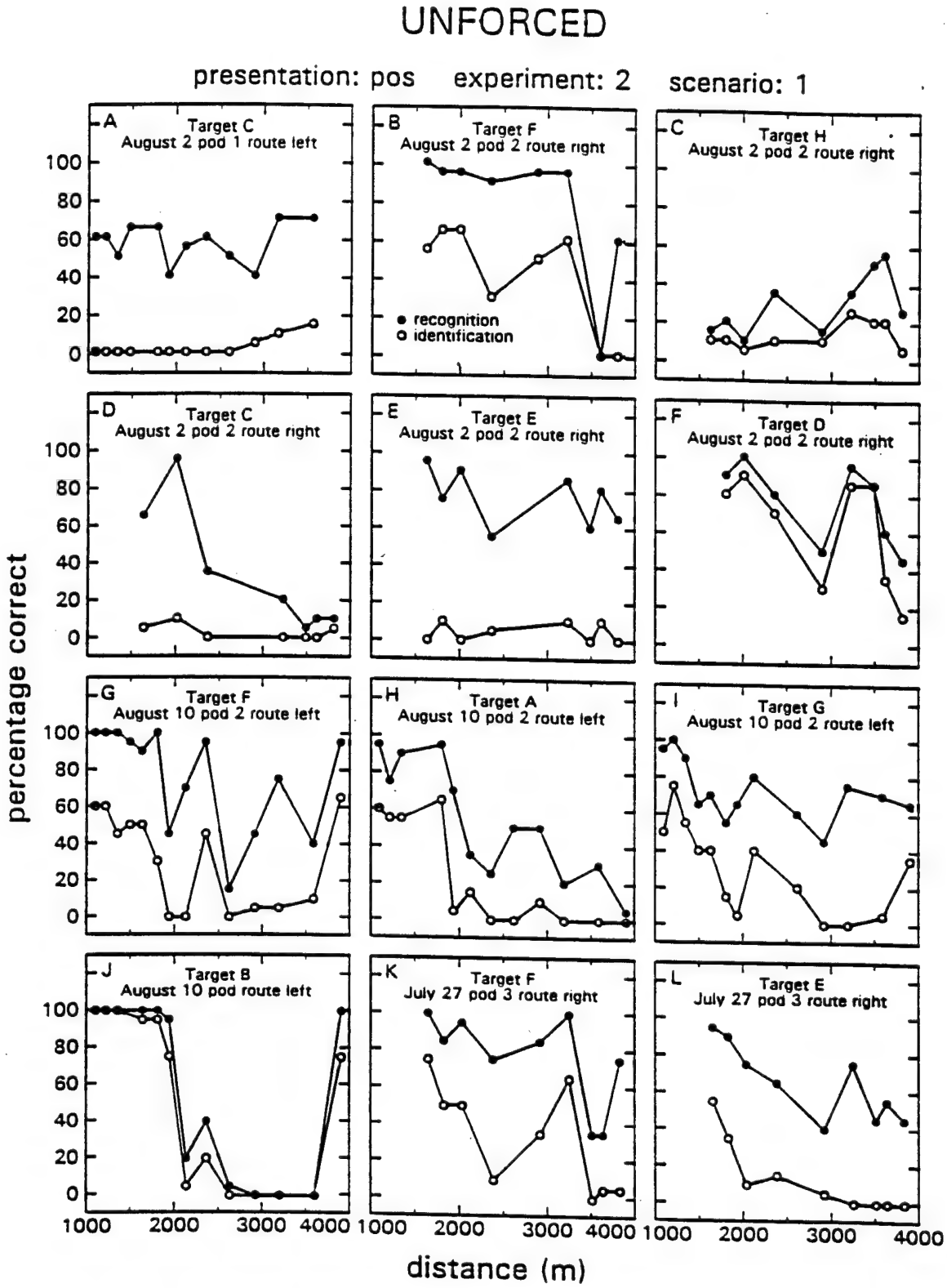


Figure A-6. Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 1, Experiment 2.

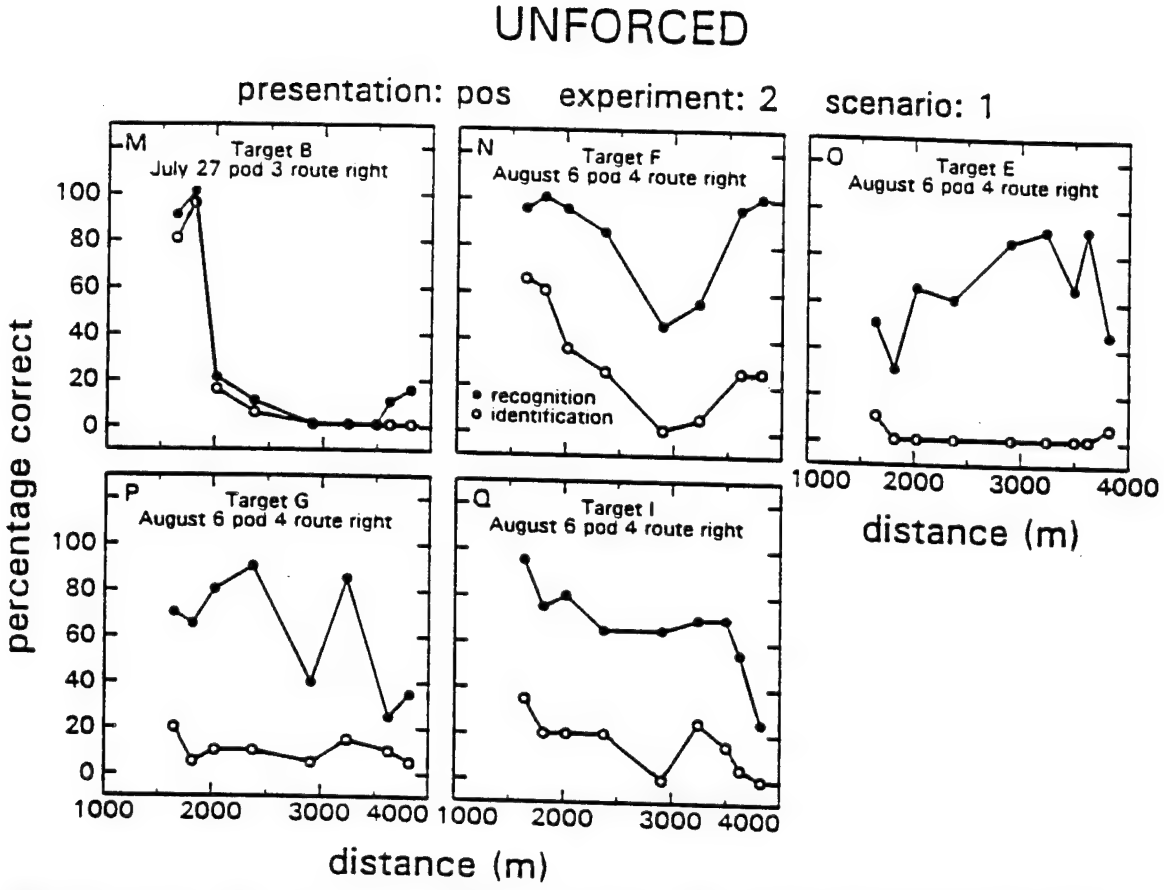


Figure A-6. Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 1, Experiment 2 (continued).

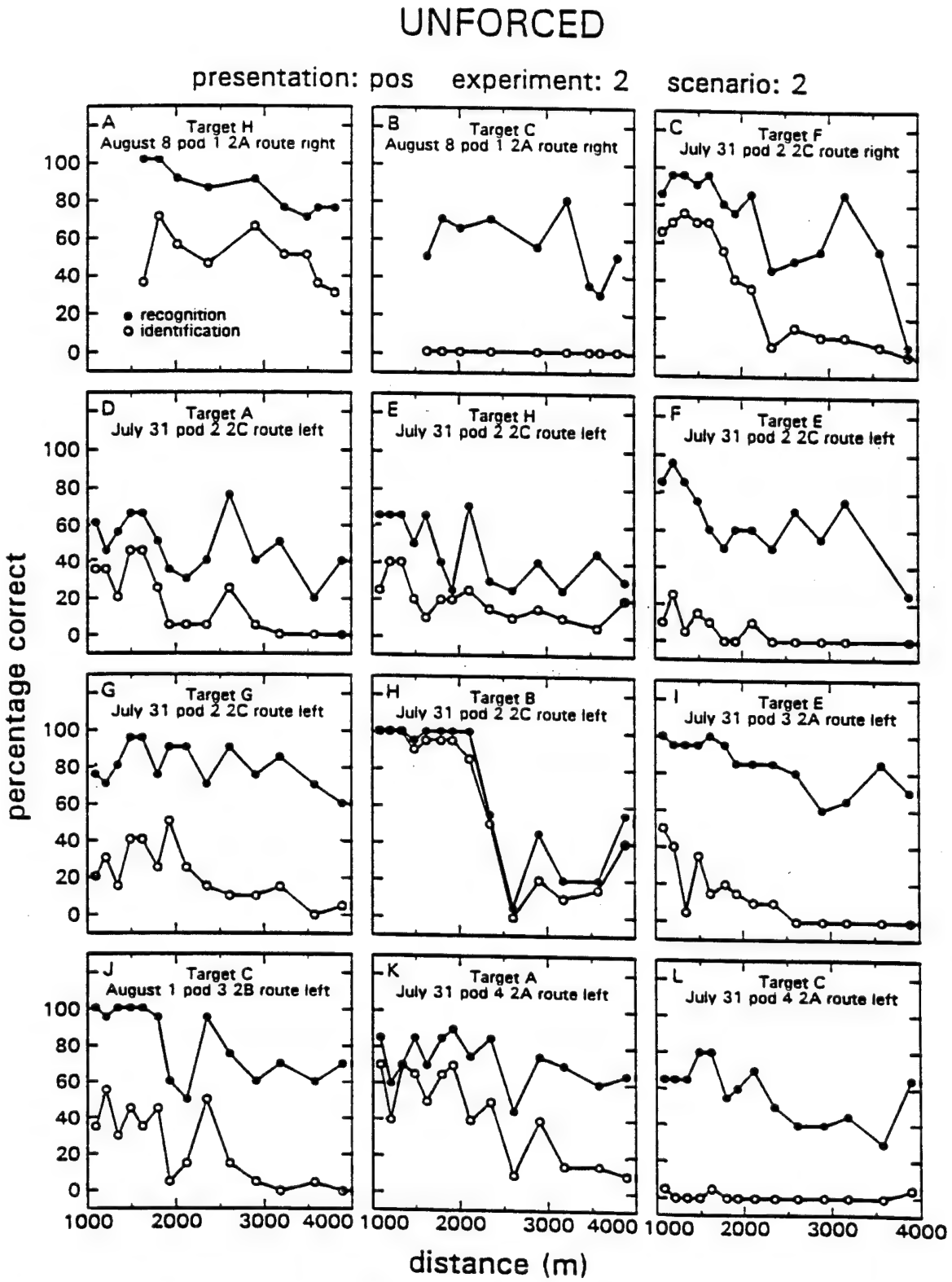


Figure A-7. Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 2, Experiment 2.

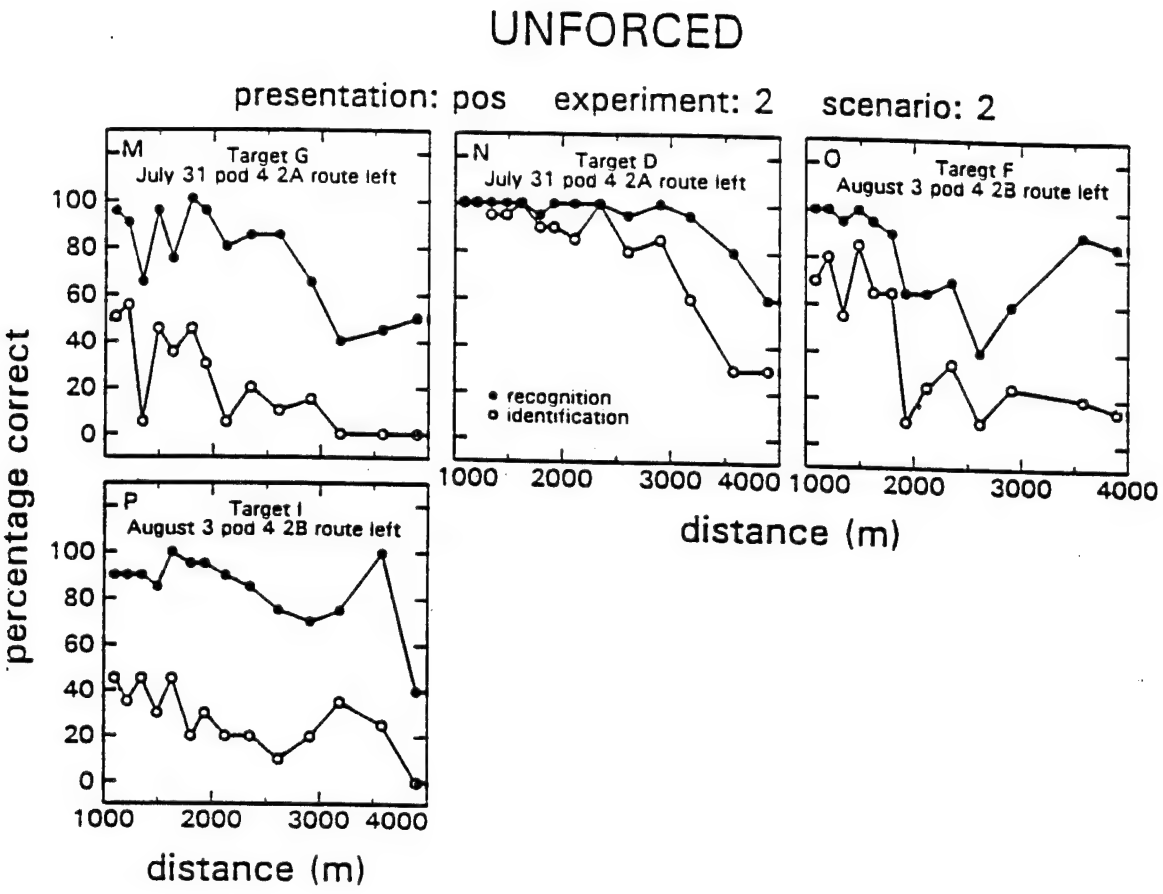


Figure A-7. Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 2, Experiment 2 (continued).

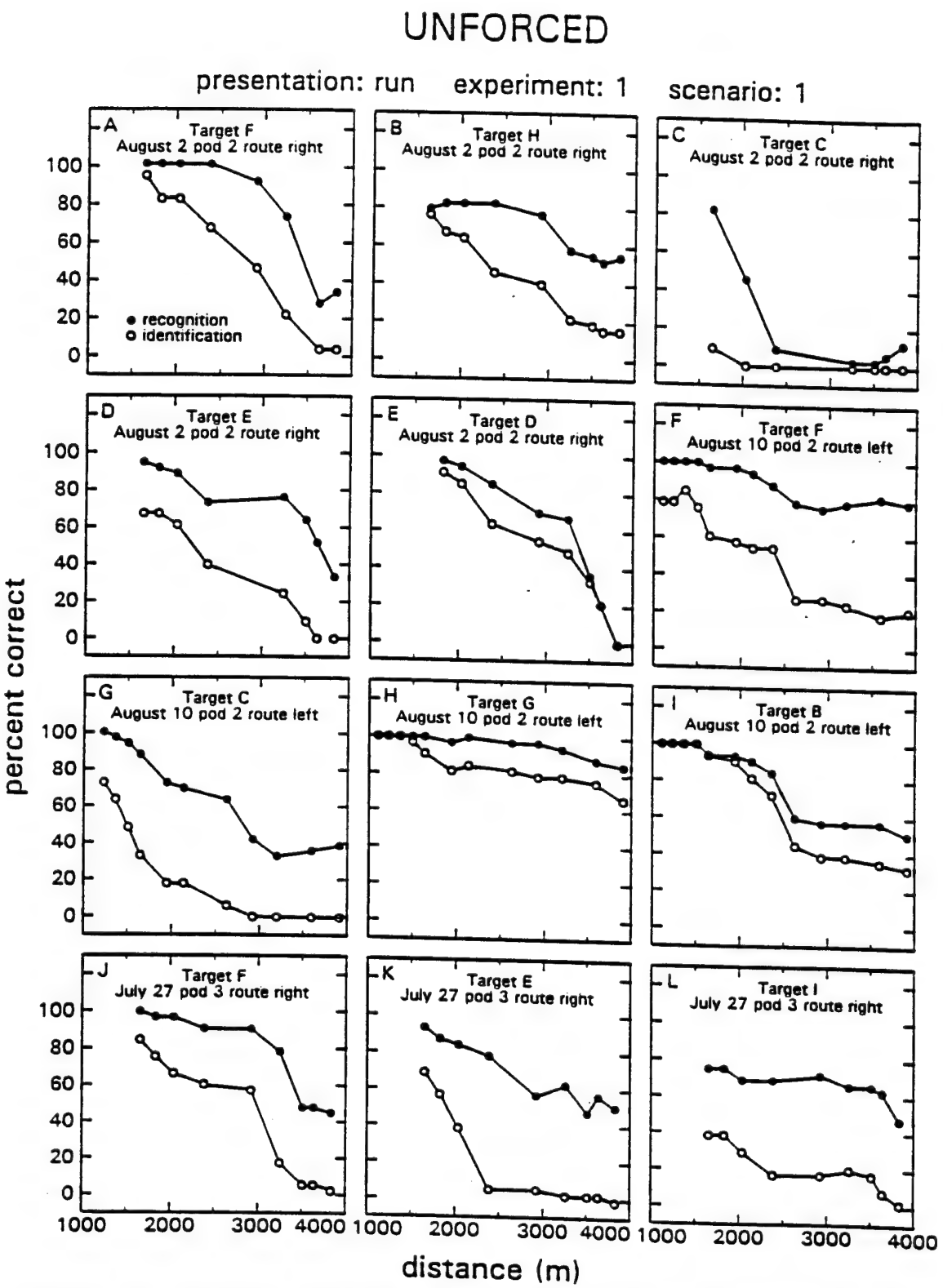


Figure A-8. Observer recognition and identification performance for the condition UNFORCED-RUN-Scenario 1, Experiment 1.

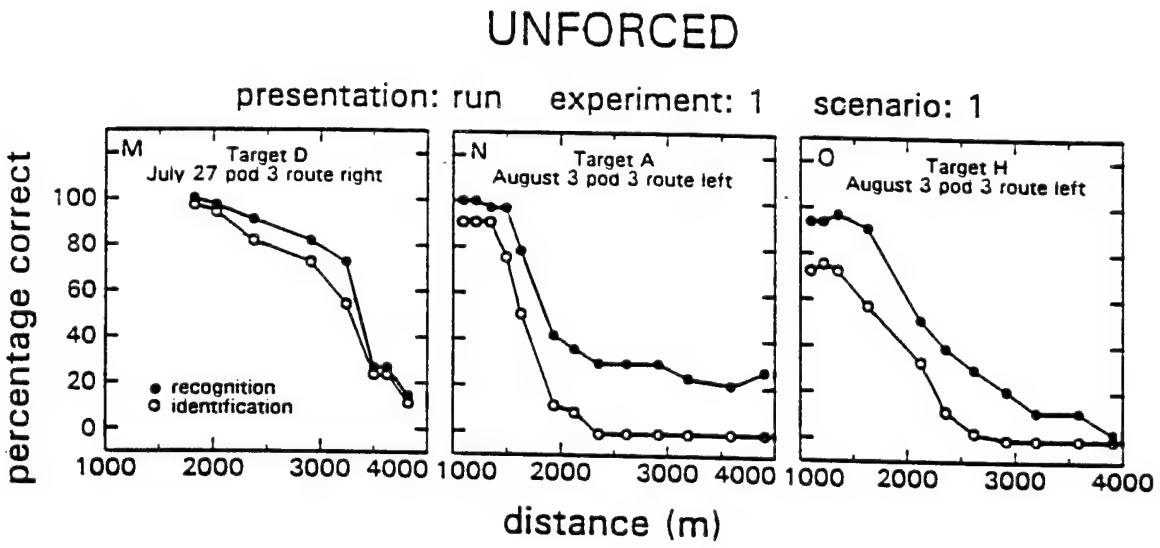


Figure A-8. Observer recognition and identification performance for the condition UNFORCED-RUN-Scenario 1, Experiment 1 (continued).

Appendix B

Recognition Performance Data with Corresponding TARGAC Predictions

The observer recognition scores are taken from figures A-5 through A-8 in appendix A. These are the scores for free unforced recognition reports, which correspond to the target acquistion (TA) task in a practical military field situation.

Figures B-1 and B-2 show the data for the position or pop-up presentation order. Figure B-4 shows the data for the sequential presentation order.

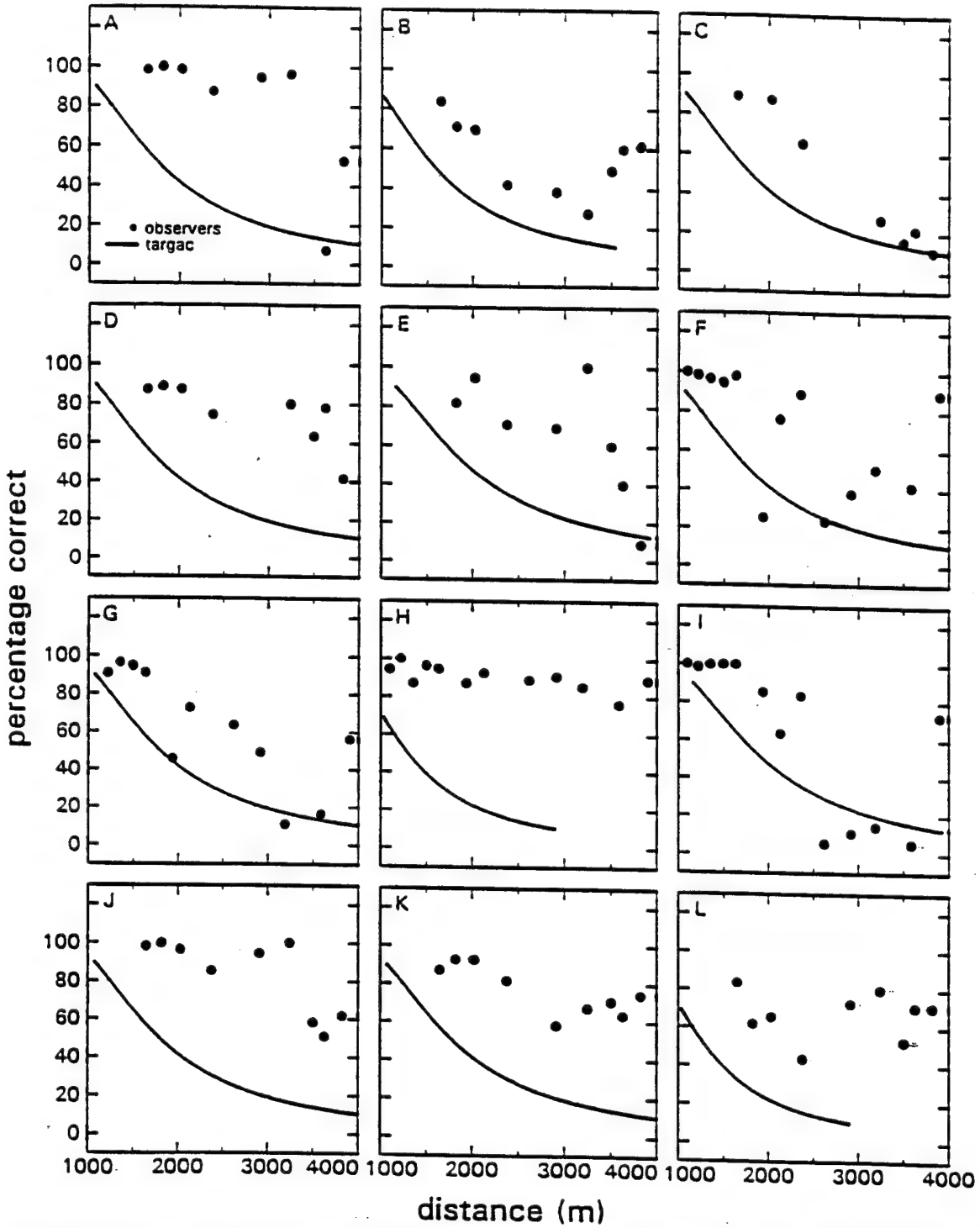


Figure B-1. Observer recognition scores for 15 runs of Experiment 1 (section 6.6.3.1) for the pop-up presentation order with TARGAC predictions.

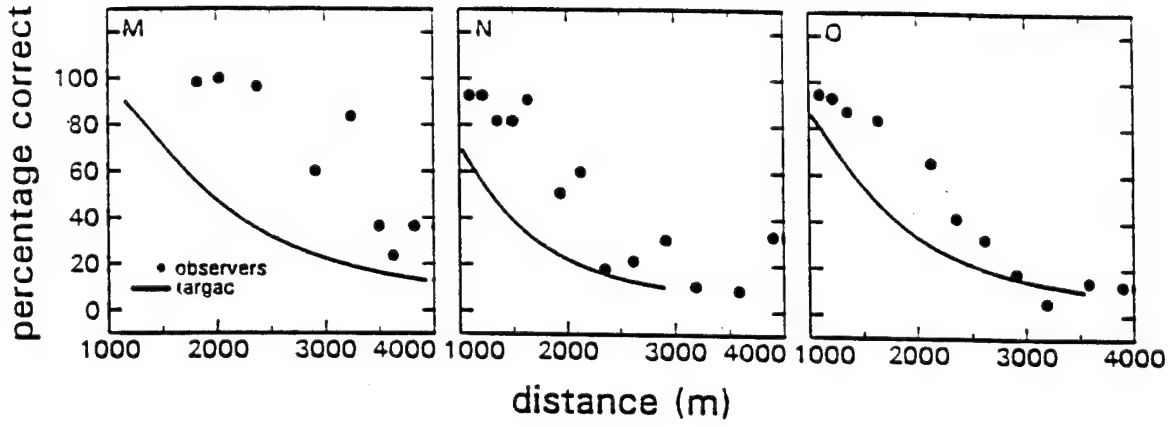


Figure B-1. Observer recognition scores for 15 runs of Experiment 1 (section 6.6.3.1) for the pop-up presentation order with TARGAC predictions (continued).

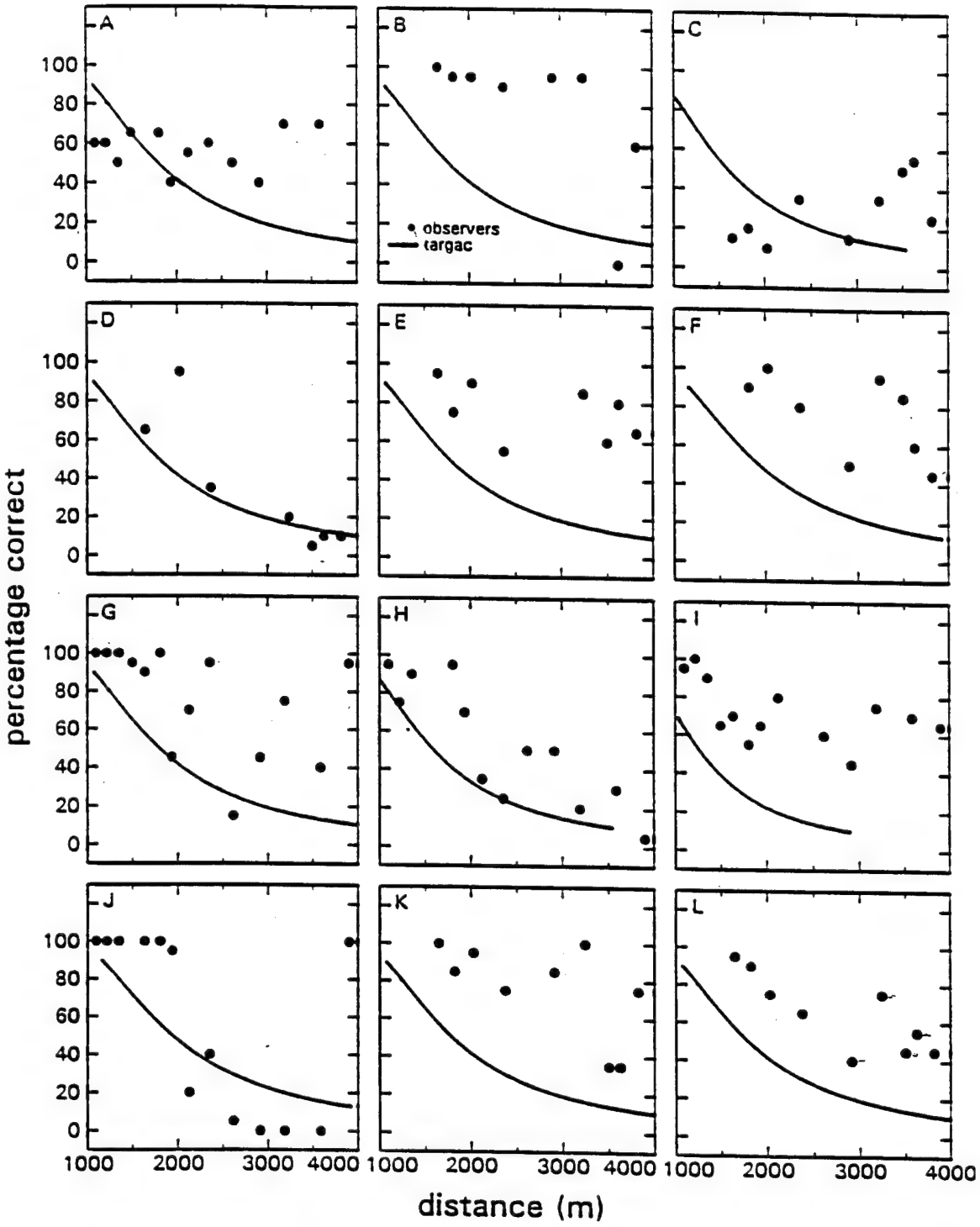


Figure B-2. Observer recognition scores for 17 runs of Experiment 2 (section 6.6.3.1) for stationary targets with TARGAC predictions.

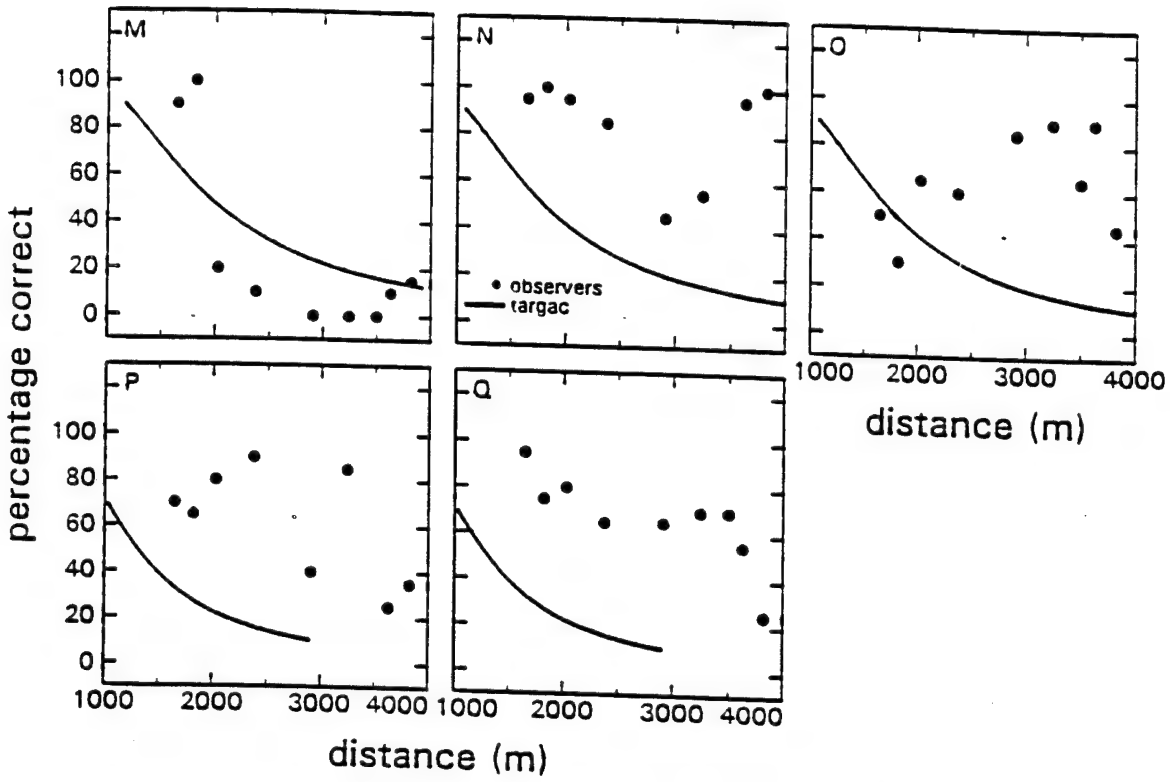


Figure B-2. Observer recognition scores for 17 runs of Experiment 2 (section 6.6.3.1) for stationary targets with TARGAC predictions (continued).

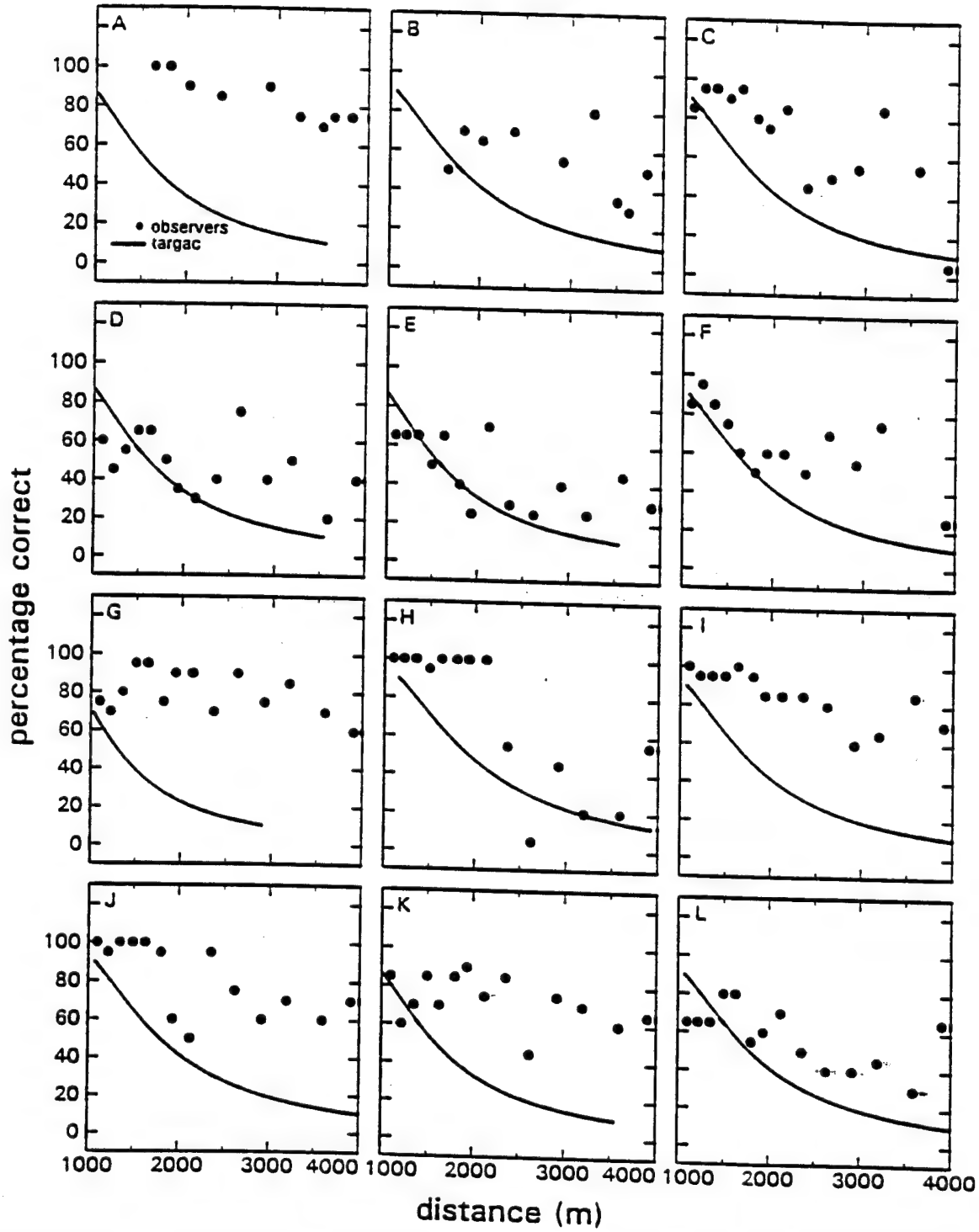


Figure B-3. Observer recognition scores for 16 runs of Experiment 2 (section 6.6.3.1) for moving targets with TARGAC predictions.

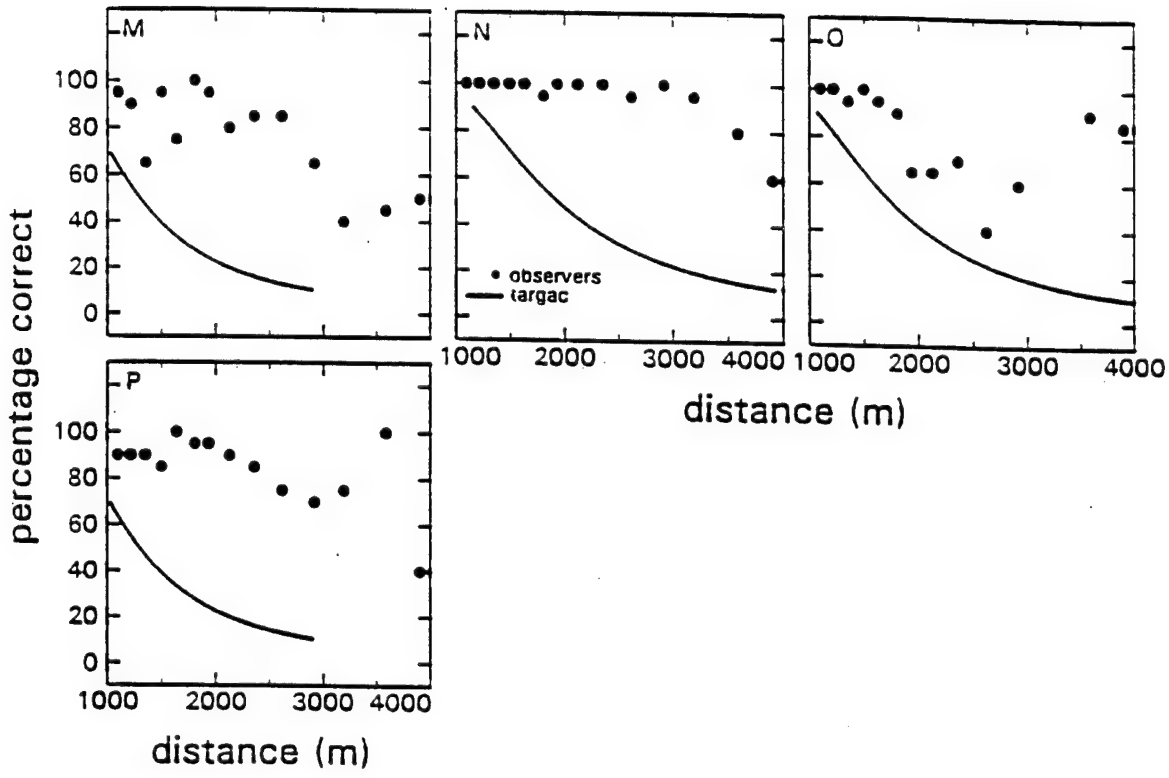


Figure B-3. Observer recognition scores for 16 runs of Experiment 2 (section 6.6.3.1) for moving targets with TARGAC predictions (continued).

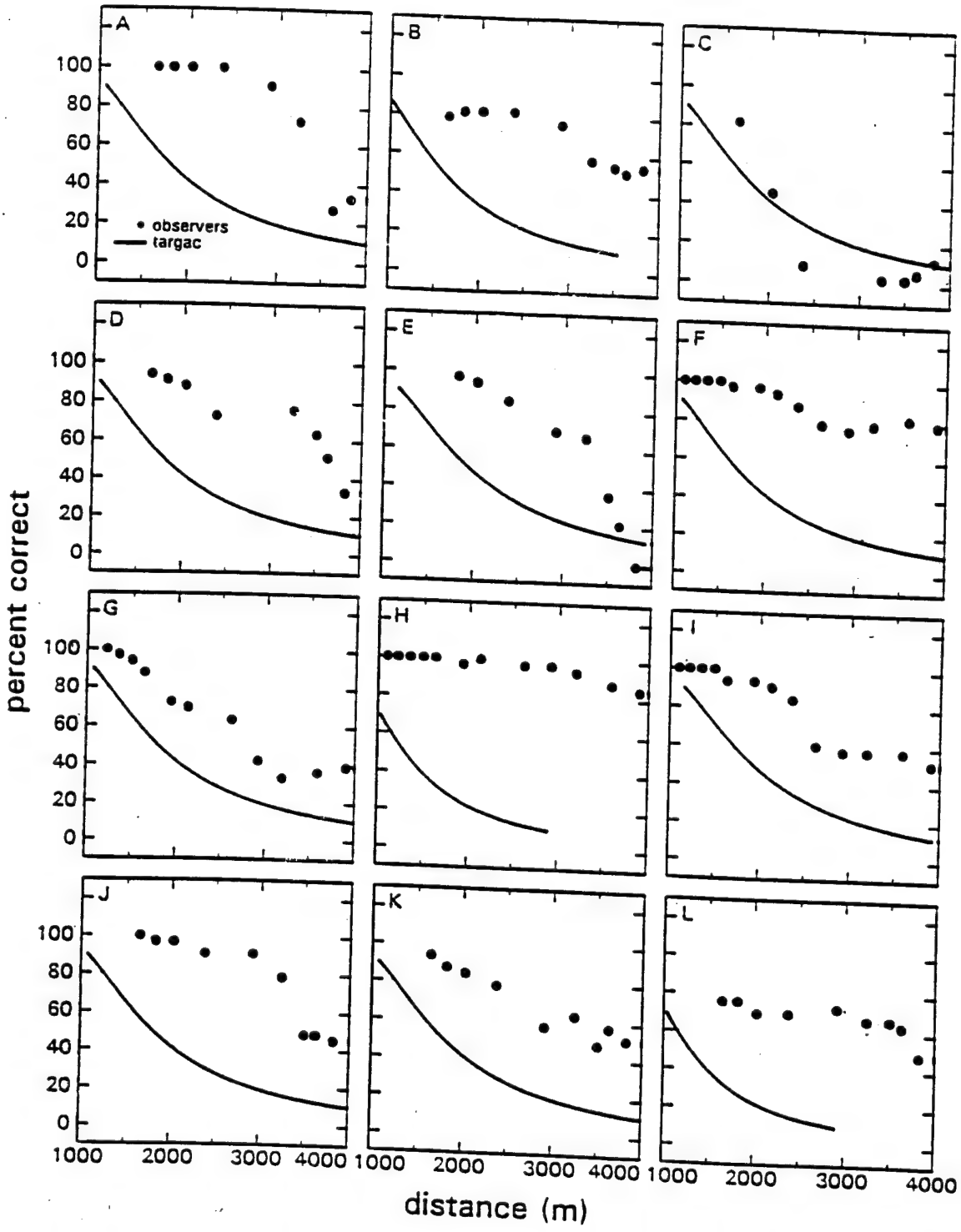


Figure B-4. Observer recognition scores for 15 runs of Experiment 1 (section 6.6.3.1) for the approaching presentation order with TARGAC predictions.

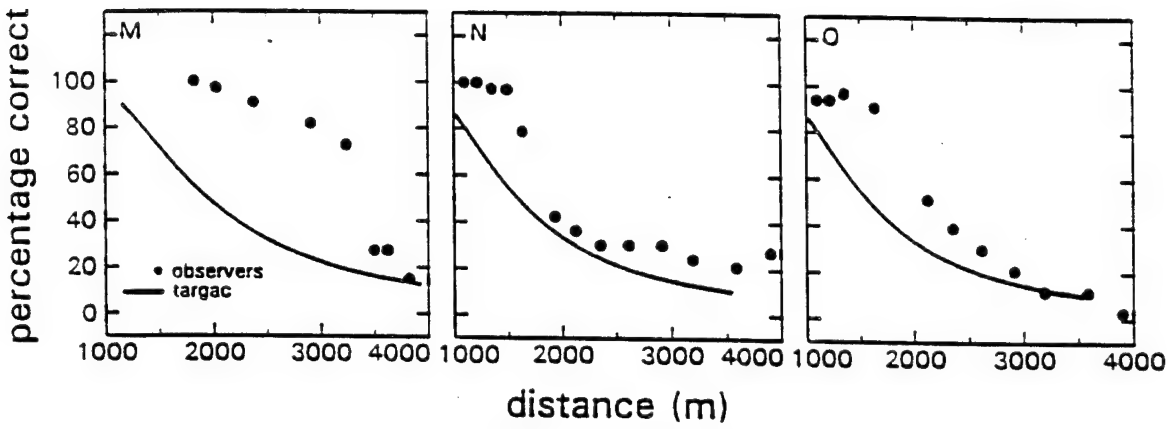


Figure B-4. Observer recognition scores for 15 runs of Experiment 1 (section 6.6.3.1) for the approaching presentation order with TARGAC predictions (continued).

Appendix C

Statistical Error in Observer Scores and Validation Accuracy

C-1. Estimate of the Error in the Observer Scores

The observer performance data were gathered in two experiments (section 4.1). The number of observers in Experiment 1 was 11. Four observers participated in Experiment 2. The images of 10 target runs were used in both experiments; thus, were presented to 15 observers. Data set A was composed of observer performance data from Experiments 1 and 2 (which means that the number of observers is 4 to 15). Data sets B and C were composed of data only from Experiment 1 (11 observers).

Each target image was presented five times. Section 3 shows that the maximum standard deviation σ_{max}, being the standard deviation at the 50 percent probability level, is 7 to 15 percent for 11 observers and 11 to 25 percent for 4 observers if the scores are distributed binomially. The lower estimates of σ_{max} are based on the assumptions that all observations are independent. A worst-case assumption was made that the five observations of the same image by the same observer are completely dependent for the higher values.

The standard deviation at probability levels above or below 50 percent is given by

$$\sigma_p = \sigma_{max} \frac{\sqrt{P\ (100 - P)}}{50} \qquad \text{(C-1)}$$

where

P = the probability level in percent.

From equation (C-1), it can be deduced that the standard deviation is rather constant at levels between 20 and 80 percent and drops to zero if the probability is near 0 or 100 percent.

The above-mentioned estimates of σ_{max} are not very accurate. The standard deviation can be estimated in an alternative way by dividing the observers into two groups and calculating the correlation between the scores of the two groups. The statistical error in the scores is small if the correlation

coefficient r is high. It can be shown that an accurate approximation of the error variance (this is the variance in the scores for each image, because of a limited number of observations), is given by

$$\sigma_e^2 \approx \frac{(\sigma_{G1}^2 + \sigma_{G2}^2)\ (1-r)}{4} \qquad \text{(C-2)}$$

where

σ_e^2 = the error variance in the score for each image, averaged over all observations of all observers of the two groups

σ^2_{G1} and σ^2_{G2} = total variance in the scores for the images, when averaged over the observations of the observers within each group.

The correlation was calculated for the images of the 10 runs that were presented to 15 observers. Observers from the two experiments were divided equally over the two groups. The correlation is very high if all data (96 datapoints) are taken into account, r = 0.92, yielding a standard deviation σ_e = 4.1 percent. However, this is an average over all probability levels between 0 and 100 percent. The standard deviation for low and high scores is much lower than for intermediate probability levels. As indicated above, for probability levels between 20 and 80 percent, the standard deviation may be regarded as rather constant. An approximation for the maximum standard deviation σ_{max} at the 50 percent level is achieved if only these data (38 datapoints) are taken into account. It turns out that for these data the correlation is still very high: r = 0.86, and σ_e = 5.0 percent for these data. This is the estimate of the maximum standard deviation for 15 observers. As the standard deviation is inversely proportional to the square-root of the number of observers, the maximum standard deviation in the data in set A (4 to 15 observers) is σ_{max} = 5 to 10 percent, and in set B and C (11 observers) is σ_{max} = 6 percent. These values are comparable to the lower estimates based on a binomial distribution.

C-2. Contribution to the Error in the (r/r′) Values

Actual target range r is known with high accuracy for each trial. The predicted range r' is calculated from the value of the recognition probability P by using the Target Acquisition model, or equation (4) in section 6.5.1. An error σ_P in probability will result in an error $\sigma_{r',obs}$ in range or $\sigma_{(r/r'),obs}$ in the ratio. The ratio ($\sigma_P / \sigma_{r',obs}$) is given by the slope of the probability versus range function (equation (C-3)).

For example, at the 50 percent level the slope of the function (with s = 2.32) is $0.80/r_{50}$. A small error in recognition probability may lead to a large error in range because the function is shallow at very high or low levels. Between 20 and 80 percent, the standard deviation is approximately constant and the probability versus range function is roughly linear. Therefore it is safe to consider only data between these probability levels. It can be shown that, for these data, the standard deviation in ratio, caused by the error in observer scores, is given by

$$\sigma_{(\frac{r}{r'}),obs} \approx 1.5\ \sigma_{max} \qquad \text{(C-3)}$$

Thus, for data set A, $\sigma_{(r/r'),obs} \approx$ 8 to 15 percent (0.03 to 0.06 on a log scale), and for sets B and C, $\sigma_{(r/r'),obs} \approx$ 9 percent (0.04 on a log scale).

Appendix D

Software Errors in TARGAC

The following software errors were found during the evaluation of the Target Acquisition Model (TARGAC).

D-1. Conversion Errors

For a single variable, several units are used in the program. For example, size or distance are expressed in meters, kilometers, or feet. It occurs in a number of subroutines that global variables, the value of which is written in a common block, are converted from one unit to another. The new value of the variable is used in the next routines through the common block, although a number of those routines expect the variable in the old unit. Incidentally, the same conversion is carried out twice. This occurs with the conversion of the rain rate in mm/h to in/h. It is recommended that the value or unit of global variables is never converted. A local variable should be defined if the value of a variable is desired in a different unit.

D-2. Sensor Altitude

Target and sensor height may be varied in TARGAC; however, a supplement of the User's Guide reports that the slant path option (looking down to a target) does not work correctly and the sensor altitude is currently hardwired to 0 or 1 m. What actually happens in the subroutine that calculates acquisition ranges (FINDR) is that sensor altitude is temporarily set to 1 and later to 0. However, ranges are defined in km in this routine. Thus, the program calculates ranges for a sensor looking down from a height of 1 km; whereas, the output file gives an altitude of 0. This error has important consequences for the range predictions. The minimum or effective dimension is target height for a target in front view. However, looking down from an altitude of 1 km, its effective dimension is target width, which is usually larger. The error was repaired by setting the altitude to 0 in FINDR.

D-3. Extrapolation of a High Order Polynomial Curve Fit

When predictions are being made for a user-specified viewing device (section 3.2), the user has to specify spatial frequency as a function of

luminance or thermal contrast in the form of the coefficients of a sixth order polynomial fit rather than a point-by-point entry of the minimum resolvable contrast (MRC) or minimum resolvable temperature difference (MRTD) curve. Point-by-point entry is another TARGAC option to specify a viewing device, but it does not work properly. Polynomial curve fits may only be used for interpolation. Extrapolation of a high-order polynomial function may lead to irrational results. However, thermal contrasts, as calculated by the thermal contrast model (TCM2), are often much higher than the highest contrast in the MRTD curve. As a consequence, unrealistic spatial frequency values are calculated which cause very high values and even negative values occur. This in return lead to meaningless range predictions. No warning is given to the user. The problem probably also occurs with viewing devices that are in the TARGAC menu. [22] The following improvements are recommended:

1. A warning should be given (actually, a lower contrast limit already exists in TARGAC) if the apparent contrast exceeds a given limit (the highest contrast of the MRC or MRTD curve).

2. Calculations may be carried out using the spatial frequency that corresponds to the contrast limit if the limit is exceeded.

3. A lower-order polynomial fit should be used; a second or third order fit should be sufficient. In the present evaluation, the problem was circumvented by adding extra MRTD points for thermal contrasts up to 20 K before making a polynomial fit.

D-4. History of Meteorological Data

TCM2 calculates target and background temperature at a given moment on the basis of meteorological data for a number of earlier moments in time, (0, 3, and 6 h earlier). The data are treated differently in interactive and batch mode. The User's Guide is correct only for the interactive mode. In batch mode, history is reversed if the input file is constructed according to the User's Guide. A correct calculation is made if the preceding times are given as negative numbers. Accordingly, files saved in the interactive mode, and input files for the batch mode are incompatible with respect to this point.

D-5. Target Heading

The user is free to choose the heading of the target in the input. However, target heading is always set at 90° in the program.

D-6. Wrong Target Files

TCM2 rendered a temperature of 0 K for several targets in the menu (targets 19 through 22). Range calculations were made using this target temperature, and no warning was given to the user. Later, new target files that gave reasonable temperature values were provided by Dr. Gillespie.

D-7. Version Number and Release Date

Regularly, new versions of the program have been released. However, the version number and release date are not updated.

D-8. Undeclared Variables

A number of variables used in the program are not declared.

Distribution

	Copies
ARMY CHEMICAL SCHOOL ATZN CM CC ATTN MR BARNES FT MCCLELLAN AL 36205-5020	1
NASA MARSHAL SPACE FLT CTR ATMOSPHERIC SCIENCES DIV E501 ATTN DR FICHTL HUNTSVILLE AL 35802	1
NASA SPACE FLT CTR ATMOSPHERIC SCIENCES DIV CODE ED 41 1 HUNTSVILLE AL 35812	1
ARMY STRAT DEFNS CMND CSSD SL L ATTN DR LILLY PO BOX 1500 HUNTSVILLE AL 35807-3801	1
ARMY MISSILE CMND AMSMI RD AC AD ATTN DR PETERSON REDSTONE ARSENAL AL 35898-5242	1
ARMY MISSILE CMND AMSMI RD AS SS ATTN MR H F ANDERSON REDSTONE ARSENAL AL 35898-5253	1

ARMY MISSILE CMND
AMSMI RD AS SS
ATTN MR B WILLIAMS
REDSTONE ARSENAL
AL 35898-5253 1

ARMY MISSILE CMND
AMSMI RD DE SE
ATTN MR GORDON LILL JR
REDSTONE ARSENAL
AL 35898-5245 1

ARMY MISSILE CMND
REDSTONE SCI INFO CTR
AMSMI RD CS R DOC
REDSTONE ARSENAL
AL 35898-5241 1

ARMY MISSILE CMND
AMSMI
REDSTONE ARSENAL
AL 35898-5253 1

ARMY INTEL CTR
AND FT HUACHUCA
ATSI CDC C
FT HUACHUCA AZ 85613-7000 1

NORTHROP CORPORATION
ELECTR SYST DIV
ATTN MRS T BROHAUGH
2301 W 120TH ST BOX 5032
HAWTHORNE CA 90251-5032 1

NAVAL WEAPONS CTR
CODE 3331
ATTN DR SHLANTA
CHINA LAKE CA 93555 1

PACIFIC MISSILE TEST CTR 1
GEOPHYSICS DIV
ATTN CODE 3250
POINT MUGU CA 93042-5000

LOCKHEED MIS & SPACE CO 1
ATTN KENNETH R HARDY
ORG 91 01 B 255
3251 HANOVER STREET
PALO ALTO CA 94304-1191

NAVAL OCEAN SYST CTR 1
CODE 54
ATTN DR RICHTER
SAN DIEGO CA 92152-5000

METEOROLOGIST IN CHARGE 1
KWAJALEIN MISSILE RANGE
PO BOX 67
APO SAN FRANCISCO
CA 96555

DEPT OF COMMERCE CTR 1
MOUNTAIN ADMINISTRATION
SPPRT CTR LIBRARY R 51
325 S BROADWAY
BOULDER CO 80303

DR HANS J LIEBE 1
NTIA ITS S 3
325 S BROADWAY
BOULDER CO 80303

NCAR LIBRARY SERIALS 1
NATL CTR FOR ATMOS RSCH
PO BOX 3000
BOULDER CO 80307-3000

DEPT OF COMMERCE CTR 1
325 S BROADWAY
BOULDER CO 80303

DAMI POI
WASH DC 20310-1067 1

MIL ASST FOR ENV SCI OFC
OF THE UNDERSEC OF DEFNS
FOR RSCH & ENGR R&AT E LS
PENTAGON ROOM 3D129
WASH DC 20301-3080 1

DEAN RMD
ATTN DR GOMEZ
WASH DC 20314 1

SPACE NAVAL WARFARE
SYST CMND
PMW 145 1G
WASH DC 20362-5100 1

ARMY INFANTRY
ATSH CD CS OR
ATTN DR E DUTOIT
FT BENNING GA 30905-5090 1

AIR WEATHER SERVICE
TECH LIBRARY FL4414 3
SCOTT AFB IL 62225-5458 1

USAFETAC DNE
ATTN MR GLAUBER
SCOTT AFB IL 62225-5008 1

HQ AWS DOO 1
SCOTT AFB IL 62225-5008 1

ARMY SPACE INSTITUTE
ATTN ATZI SI 3
FT LEAVENWORTH
KS 66027-5300 1

PHILLIPS LABORATORY 1
PL LYP
ATTN MR CHISHOLM
HANSCOM AFB MA 01731-5000

ATMOSPHERIC SCI DIV 1
GEOPHYSICS DIRCTRT
PHILLIPS LABORATORY
HANSCOM AFB MA 01731-5000

PHILLIPS LABORATORY 1
PL LYP 3
HANSCOM AFB MA 01731-5000

RAYTHEON COMPANY 1
ATTN DR SONNENSCHEIN
528 BOSTON POST ROAD
SUDBURY MA 01776
MAIL STOP 1K9

ARMY MATERIEL SYST 1
ANALYSIS ACTIVITY
AMXSY
ATTN MP H COHEN
APG MD 21005-5071

ARMY MATERIEL SYST 1
ANALYSIS ACTIVITY
AMXSY AT
ATTN MR CAMPBELL
APG MD 21005-5071

ARMY MATERIEL SYST 1
ANALYSIS ACTIVITY
AMXSY CR
ATTN MR MARCHET
APG MD 21005-5071

ARL CHEMICAL BIOLOGY 1
NUC EFFECTS DIV
AMSRL SL CO
APG MD 21010-5423

ARMY MATERIEL SYST 1
ANALYSIS ACTIVITY
AMXSY
APG MD 21005-5071

NAVAL RESEARCH LABORATORY 1
CODE 4110
ATTN MR RUHNKE
WASH DC 20375-5000

ARMY MATERIEL SYST 1
ANALYSIS ACTIVITY
AMXSY CS
ATTN MR BRADLEY
APG MD 21005-5071

ARMY RESEARCH LABORATORY 1
AMSRL D
2800 POWDER MILL ROAD
ADELPHI MD 20783-1145

ARMY RESEARCH LABORATORY 1
AMSRL OP SD TP
TECHNICAL PUBLISHING
2800 POWDER MILL ROAD
ADELPHI MD 20783-1145

ARMY RESEARCH LABORATORY 1
AMSRL OP CI SD TL
2800 POWDER MILL ROAD
ADELPHI MD 20783-1145

ARMY RESEARCH LABORATORY 1
AMSRL SS SH
ATTN DR SZTANKAY
2800 POWDER MILL ROAD
ADELPHI MD 20783-1145

ARMY RESEARCH LABORATORY 1
AMSRL
2800 POWDER MILL ROAD
ADELPHI MD 20783-1145

NATIONAL SECURITY AGCY W21 1
ATTN DR LONGBOTHUM
9800 SAVAGE ROAD
FT GEORGE G MEADE
MD 20755-6000

ARMY AVIATION CTR 1
ATZQ D MA
ATTN MR HEATH
FT RUCKER AL 36362

OIC NAVSWC 1
TECH LIBRARY CODE E 232
SILVER SPRINGS
MD 20903-5000

ARMY RSRC OFC 1
ATTN DRXRO GS
PO BOX 12211
RTP NC 27009

DR JERRY DAVIS 1
NCSU
PO BOX 8208
RALEIGH NC 27650-8208

ARMY CCREL 1
CECRL GP
ATTN DR DETSCH
HANOVER NH 03755-1290

ARMY ARDEC
SMCAR IMI I BLDG 59
DOVER NJ 07806-5000 1

ARMY SATELLITE COMM AGCY
DRCPM SC 3
FT MONMOUTH NJ 07703-5303 1

ARMY COMMUNICATIONS
ELECTR CTR FOR EW RSTA
AMSEL EW D
FT MONMOUTH NJ 07703-5303 1

ARMY COMMUNICATIONS
ELECTR CTR FOR EW RSTA
AMSEL EW MD
FT MONMOUTH NJ 07703-5303 1

ARMY DUGWAY PROVING GRD
STEDP MT DA L 3
DUGWAY UT 84022-5000 1

ARMY DUGWAY PROVING GRD
STEDP MT M
ATTN MR BOWERS
DUGWAY UT 84022-5000 1

DEPT OF THE AIR FORCE
OL A 2D WEATHER SQUAD MAC
HOLLOMAN AFB
NM 88330-5000 1

PL WE
KIRTLAND AFB NM
87118-6008 1

USAF ROME LAB TECH
CORRIDOR W STE 262 RL SUL
26 ELECTR PKWY BLD 106
GRIFFISS AFB
NY 13441-4514 1

AFMC DOW 1
WRIGHT PATTERSON AFB
OH 0334-5000

ARMY FIELD ARTLLRY SCHOOL 1
ATSF TSM TA
FT SILL OK 73503-5600

NAVAL AIR DEV CTR 1
CODE 5012
ATTN AL SALIK
WARMINISTER PA 18974

ARMY FOREGN SCI TECH CTR 1
CM
220 7TH STREET NE
CHARLOTTESVILLE
VA 22901-5396

NAVAL SURFACE WEAPONS CTR 1
CODE G63
DAHLGREN VA 22448-5000

ARMY OEC 1
CSTE EFS
PARK CENTER IV
4501 FORD AVE
ALEXANDRIA VA 22302-1458

ARMY CORPS OF ENGRS 1
ENGR TOPOGRAPHICS LAB
ETL GS LB
FT BELVOIR VA 22060

TAC DOWP 1
LANGLEY AFB
VA 23665-5524

ARMY TOPO ENGR CTR 1
CETEC ZC 1
FT BELVOIR VA 22060-5546

LOGISTICS CTR
ATCL CE
FT LEE VA 23801-6000 1

SCI AND TECHNOLOGY
101 RESEARCH DRIVE
HAMPTON VA 23666-1340 1

ARMY NUCLEAR CML AGCY
MONA ZB BLDG 2073
SPRINGFIELD VA 22150-3198 1

ARMY FIELD ARTLLRY SCHOOL
ATSF F FD
FT SILL OK 73503-5600 1

USATRADOC
ATCD FA
FT MONROE VA 23651-5170 1

ARMY TRADOC ANALYSIS CTR
ATRC WSS R
WSMR NM 88002-5502 1

ARMY RESEARCH LABORATORY
AMSRL BE M
BATTLEFIELD ENVIR DIR
WSMR NM 88002-5501 1

ARMY RESEARCH LABORATORY
AMSRL BE A
BATTLEFIELD ENVIR DIR
WSMR NM 88002-5501 1

ARMY RESEARCH LABORATORY
AMSRL BE W
BATTLEFIELD ENVIR DIR
WSMR NM 88002-5501 1

ARMY RESEARCH LABORATORY AMSRL BE ATTN MR VEAZEY BATTLEFIELD ENVIR DIR WSMR NM 88002-5501	1
DEFNS TECH INFO CTR CENTER DTIC BLS BLDG 5 CAMERON STATION ALEXANDRIA VA 22304-6145	1
ARMY MISSILE CMND AMSMI REDSTONE ARSENAL AL 35898-5243	1
ARMY DUGWAY PROVING GRD STEDP 3 DUGWAY UT 84022-5000	1
USATRADOC ATCD FA FT MONROE VA 23651-5170	1
ARMY FIELD ARTLRY SCHOOL ATSF FT SILL OK 73503-5600	1
WSMR TECH LIBRARY BR STEWS IM IT WSMR NM 88001	1
Record Copy	10
Total	96